Preface
This module introduces the Connexions online textbook "Collaborative Statistics" by Barbara Illowsky and Susan Dean.

Welcome to *Collaborative Statistics*, presented by Connexions. The initial section below introduces you to Connexions. If you are familiar with Connexions, please skip to About "Collaborative Statistics."

About Connexions

Connexions Modular Content
Connexions (cnx.org) is an online, **open access** educational resource dedicated to providing high quality learning materials free online, free in printable PDF format, and at low cost in bound volumes through print-on-demand publishing. The *Collaborative Statistics* textbook is one of many **collections** available to Connexions users. Each **collection** is composed of a number of re-usable learning **modules** written in the Connexions XML markup language. Each module may also be re-used (or 're-purposed') as part of other collections and may be used outside of Connexions. Including *Collaborative Statistics*, Connexions currently offers over 6500 modules and more than 350 collections.

The modules of *Collaborative Statistics* are derived from the original paper version of the textbook under the same title, *Collaborative Statistics*. Each module represents a self-contained concept from the original work. Together, the modules comprise the original textbook.

Re-use and Customization
The Creative Commons (CC) Attribution license applies to all Connexions modules. Under this license, any module in Connexions may be used or modified for any purpose as long as proper attribution to the original author(s) is maintained. Connexions' authoring tools make re-use (or re-purposing) easy. Therefore, instructors anywhere are permitted to create customized versions of the *Collaborative Statistics* textbook by editing modules, deleting unneeded modules, and adding their own supplementary modules. Connexions' authoring tools keep track of these changes and maintain the CC license's required attribution to the original authors. This

process creates a new collection that can be viewed online, downloaded as a single PDF file, or ordered in any quantity by instructors and students as a low-cost printed textbook. To start building custom collections, please visit the help page, "Create a Collection with Existing Modules". For a guide to authoring modules, please look at the help page, "Create a Module in Minutes".

Read the book online, print the PDF, or buy a copy of the book.
To browse the *Collaborative Statistics* textbook online, visit the collection home page at cnx.org/content/col10522/latest. You will then have three options.

1. You may obtain a PDF of the entire textbook to print or view offline by clicking on the "Download PDF" link in the "Content Actions" box.
2. You may order a bound copy of the collection by clicking on the "Order Printed Copy" button.
3. You may view the collection modules online by clicking on the "Start >>" link, which takes you to the first module in the collection. You can then navigate through the subsequent modules by using their "Next >>" and "Previous >>" links to move forward and backward in the collection. You can jump to any module in the collection by clicking on that module's title in the "Collection Contents" box on the left side of the window. If these contents are hidden, make them visible by clicking on "[show table of contents]".

Accessibility and Section 508 Compliance

- For information on general Connexions accessibility features, please visit http://cnx.org/content/m17212/latest/.
- For information on accessibility features specific to the Collaborative Statistics textbook, please visit http://cnx.org/content/m17211/latest/.

Version Change History and Errata

- For a list of modifications, updates, and corrections, please visit http://cnx.org/content/m17360/latest/.

Adoption and Usage

- The Collaborative Statistics collection has been adopted and customized by a number of professors and educators for use in their classes. For a list of known versions and adopters, please visit http://cnx.org/content/m18261/latest/.

About "Collaborative Statistics"

Collaborative Statistics was written by Barbara Illowsky and Susan Dean, faculty members at De Anza College in Cupertino, California. The textbook was developed over several years and has been used in regular and honors-level classroom settings and in distance learning classes. Courses using this textbook have been articulated by the University of California for transfer of credit. The textbook contains full materials for course offerings, including expository text, examples, labs, homework, and projects. A Teacher's Guide is currently available in print form and on the Connexions site at http://cnx.org/content/col10547/latest/, and supplemental course materials including additional problem sets and video lectures are available at http://cnx.org/content/col10586/latest/. The on-line text for each of these collections collections will meet the Section 508 standards for accessibility.

An on-line course based on the textbook was also developed by Illowsky and Dean. It has won an award as the best on-line California community college course. The on-line course will be available at a later date as a collection in Connexions, and each lesson in the on-line course will be linked to the on-line textbook chapter. The on-line course will include, in addition to expository text and examples, videos of course lectures in captioned and non-captioned format.

The original preface to the book as written by professors Illowsky and Dean, now follows:

This book is intended for introductory statistics courses being taken by students at two– and four–year colleges who are majoring in fields other than math or engineering. Intermediate algebra is the only prerequisite. The book focuses on applications of statistical knowledge rather than the theory

behind it. The text is named *Collaborative Statistics* because students learn best by **doing**. In fact, they learn best by working in small groups. The old saying “two heads are better than one” truly applies here.

Our emphasis in this text is on four main concepts:

- thinking statistically
- incorporating technology
- working collaboratively
- writing thoughtfully

These concepts are integral to our course. Students learn the best by actively participating, not by just watching and listening. Teaching should be highly interactive. Students need to be thoroughly engaged in the learning process in order to make sense of statistical concepts. *Collaborative Statistics* provides techniques for students to write across the curriculum, to collaborate with their peers, to think statistically, and to incorporate technology.

This book takes students step by step. The text is interactive. Therefore, students can immediately apply what they read. Once students have completed the process of problem solving, they can tackle interesting and challenging problems relevant to today’s world. The problems require the students to apply their newly found skills. In addition, technology (TI-83 graphing calculators are highlighted) is incorporated throughout the text and the problems, as well as in the special group activities and projects. The book also contains labs that use real data and practices that lead students step by step through the problem solving process.

At De Anza, along with hundreds of other colleges across the country, the college audience involves a large number of ESL students as well as students from many disciplines. The ESL students, as well as the non-ESL students, have been especially appreciative of this text. They find it extremely readable and understandable. *Collaborative Statistics* has been used in classes that range from 20 to 120 students, and in regular, honor, and distance learning classes.

Susan Dean

Barbara Illowsky

Sampling and Data
This module provides a brief introduction to the field of statistics, including examples of how these topics shows up in a variety of real-life examples.

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Create and interpret frequency tables.

Introduction

You are probably asking yourself the question, "When and where will I use statistics?". If you read any newspaper or watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a news program on television, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques to analyze the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data are.

Statistics

This module introduces the concept of statistics, specifically the ability to use statistics to describe data (descriptive statistics) as well as draw conclusions (inferential statistics). An optional classroom exercise is included.

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives.

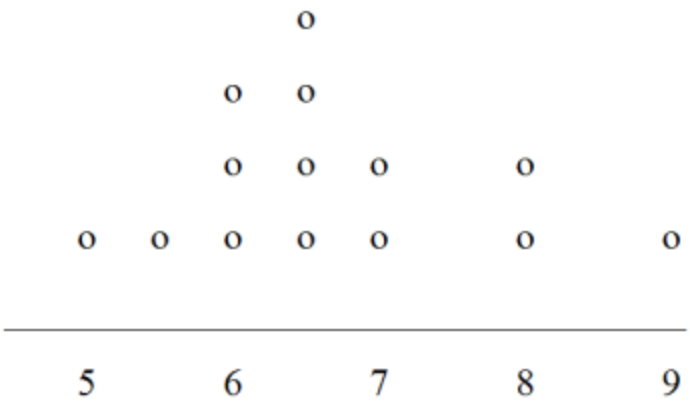
Optional Collaborative Classroom Exercise

In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:

5 5.5 6 6 6 6.5 6.5 6.5 6.5 7 7 8 8 9

The dot plot for this data would be as follows:

Frequency of Average Time (in Hours) Spent Sleeping per Night



Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

Where do your data appear to cluster? How could you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that the conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Levels of Measurement and Statistical Operations

The way a set of data is measured is called its level of measurement. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- Nominal scale level
- Ordinal scale level
- Interval scale level
- Ratio scale level

Data that is measured using a **nominal scale** is qualitative. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. Some examples are Sony, Motorola, Nokia, Samsung and Apple. This is just a list and there is no agreed upon order. Some people may favor Apple but that is a matter of opinion. Nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data but there is a big difference. The ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.

Another example using the ordinal scale is a cruise survey where the responses to questions about the cruise are “excellent,” “good,” “satisfactory” and “unsatisfactory.” These responses are ordered from the most desired response by the cruise lines to the least desired. But the differences between two pieces of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40 degrees is equal to 100 degrees minus 60 degrees. Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0.

Interval level data can be used in calculations but one type of comparison cannot be done. Eighty degrees C is not 4 times as hot as 20° C (nor is 80° F 4 times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or 4 to 1).

Data that is measured using the **ratio scale** takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data but, in addition, it has a 0 point and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams were machine-graded.

The data can be put in order from lowest to highest: 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points.

Ratios can be calculated. The smallest score for ratio data is 0. So 80 is 4 times 20. The score of 80 is 4 times better than the score of 20.

Exercises

What type of measure scale is being used? Nominal, Ordinal, Interval or Ratio.

1. High school men soccer players classified by their athletic ability: Superior, Average, Above average.
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
3. The colors of crayons in a 24-crayon box.
4. Social security numbers.
5. Incomes measured in dollars
6. A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied.
7. Political outlook: extreme left, left-of-center, right-of-center, extreme right.
8. Time of day on an analog watch.
9. The distance in miles to the closest grocery store.
10. The dates 1066, 1492, 1644, 1947, 1944.

11. The heights of 21 – 65 year-old women.
12. Common letter grades A, B, C, D, F.

Answers 1. ordinal, 2. interval, 3. nominal, 4. nominal, 5. ratio, 6. ordinal, 7. nominal, 8. interval, 9. ratio, 10. interval, 11. ratio, 12. ordinal

Glossary

Data

A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Statistic

A numerical characteristic of the sample. A statistic estimates the corresponding population parameter. For example, the average number of full-time students in a 7:30 a.m. class for this term (statistic) is an estimate for the average number of full-time students in any class this term (parameter).

Probability
This module introduces the concept of probability as a mathematical measure of randomness, including a number of real-world applications.

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a **fair** coin 4 times, the outcomes may not be 2 heads and 2 tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl Pearson who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Glossary

Probability
: A number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur. The foundation of statistics is given by the following 3 axioms (by A. N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S . Then:

  - $0 \leq P(A) \leq 1$;.
  - If A and B are any two mutually exclusive events, then $P(A \text{ or } B) = P(A) + P(B)$.

- $P(S) = 1$.

Key Terms
This module introduces a number of key terms related to statistical sampling and data.

In statistics, we generally want to study a **population**. You can think of a population as an entire collection of persons, things, or objects under study. To study the larger population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000 to 2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that is a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, notated by capital letters like X and Y, is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. Datum is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtained scores of 86, 75, and 92, you calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

Note:
Mean and Average
The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean" and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example:

Exercise:

Problem:

Define the key terms from the following study: We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly survey 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Solution:

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term.

The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

Optional Collaborative Classroom Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

Glossary

Average
: A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Data
: A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Proportion

- As a number: A proportion is the number of successes divided by the total number in the sample.
- As a probability distribution: Given a binomial random variable (RV), $X \sim B(n, p)$, consider the ratio of the number X of successes in n Bernouli trials to the number n of trials. $P\prime = \frac{X}{n}$. This new RV is called a proportion, and if the number of trials, n, is large enough, P' $\sim N\left(p, \frac{\text{pq}}{n}\right)$.

Data

This module introduces the concepts of qualitative data, quantitative continuous data, and quantitative discrete data as used in statistics. Sample problems are included.

Data may come from a population or from a sample. Small letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and the number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get 0, 1, 2, 3, etc.

All data that are the result of measuring are **quantitative continuous data** assuming that we can measure accurately. Measuring angles in radians might result in the numbers $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π, $\frac{3\pi}{4}$, etc. If you and your friends carry backpacks with books in them to school, the numbers of books in the

backpacks are discrete data and the weights of the backpacks are continuous data.

Note: In this course, the data used is mainly quantitative. It is easy to calculate statistics (like the mean or proportion) from numbers. In the chapter **Descriptive Statistics**, you will be introduced to stem plots, histograms and box plots all of which display quantitative data. Qualitative data is discussed at the end of this section through graphs.

Example:
Data Sample of Quantitative Discrete Data
The data are the number of books students carry in their backpacks. You sample five students. Two students carry 3 books, one student carries 4 books, one student carries 2 books, and one student carries 1 book. The numbers of books (3, 4, 2, and 1) are the quantitative discrete data.

Example:
Data Sample of Quantitative Continuous Data
The data are the weights of the backpacks with the books in it. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data because weights are measured.

Example:
Data Sample of Qualitative Data
The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

Note: You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

Example:
Exercise:

Problem:

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

1. The number of pairs of shoes you own.
2. The type of car you drive.
3. Where you go on vacation.
4. The distance it is from your home to the nearest grocery store.
5. The number of classes you take per school year.
6. The tuition for your classes
7. The type of calculator you use.
8. Movie ratings.
9. Political party preferences.
10. Weight of sumo wrestlers.
11. Amount of money won playing poker.
12. Number of correct answers on a quiz.
13. Peoples' attitudes toward the government.
14. IQ scores. (This may cause some discussion.)

Solution:

Items 1, 5, 11, and 12 are quantitative discrete; items 4, 6, 10, and 14 are quantitative continuous; and items 2, 3, 7, 8, 9, and 13 are qualitative.

Qualitative Data Discussion

Below are tables of part-time vs full-time students at De Anza College in Cupertino, CA and Foothill College in Los Altos, CA for the Spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

	Number	Percent
Full-time	9,200	40.9%
Part-time	13,296	59.1%
Total	22,496	100%

De Anza College

	Number	Percent
Full-time	4,059	28.6%
Part-time	10,124	71.4%

Total	14,183	100%

Foothill College

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning what graphs to use. Below are pie charts and bar graphs, two graphs that are used to display qualitative data.

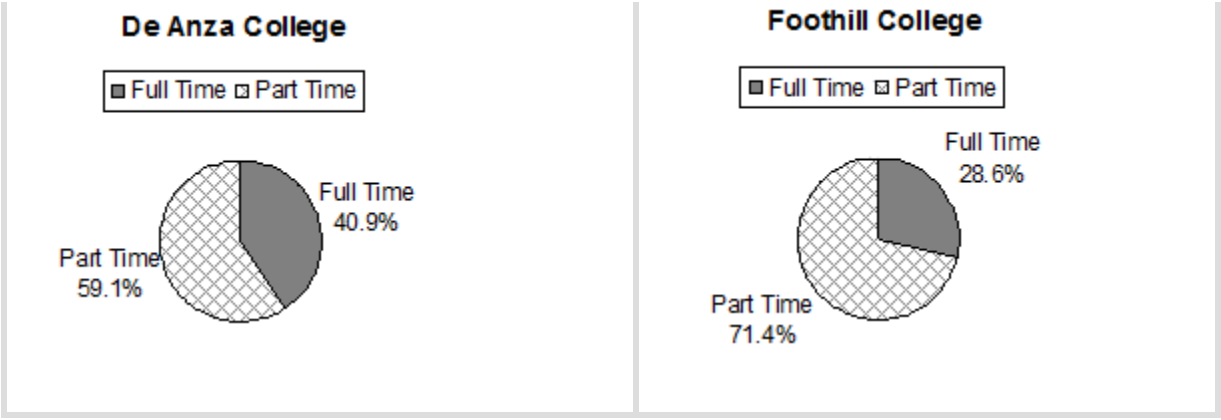
In a **pie chart**, categories of data are represented by wedges in the circle and are proportional in size to the percent of individuals in each category.

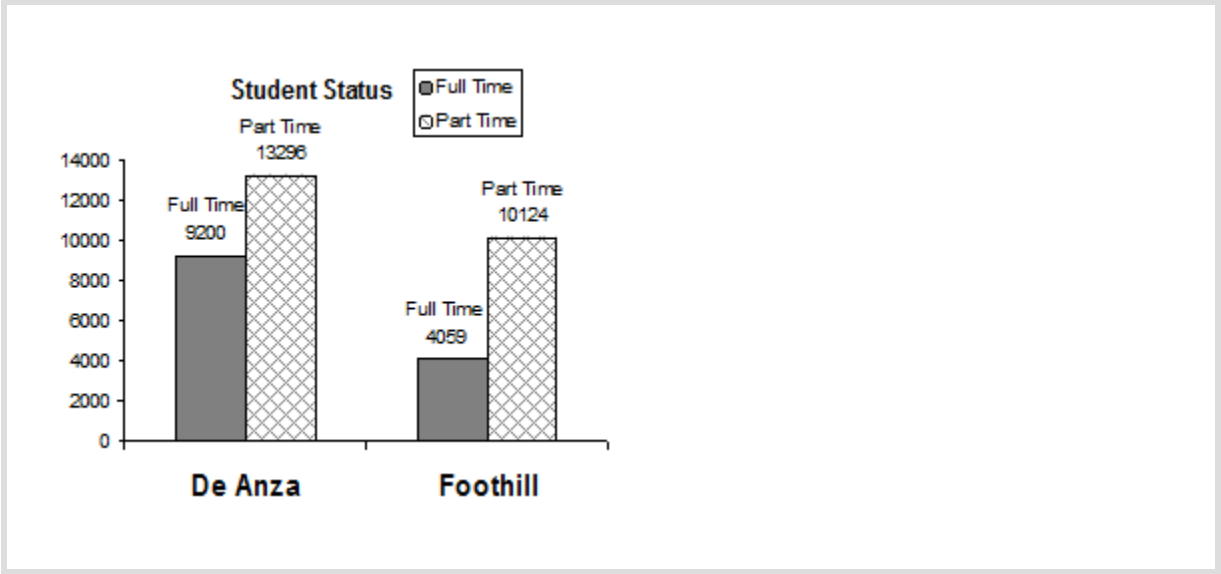
In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.

A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at the graphs and determine which graph (pie or bar) you think displays the comparisons better. This is a matter of preference.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the "best" graph depending on the data and the context. Our choice also depends on what we are using the data for.



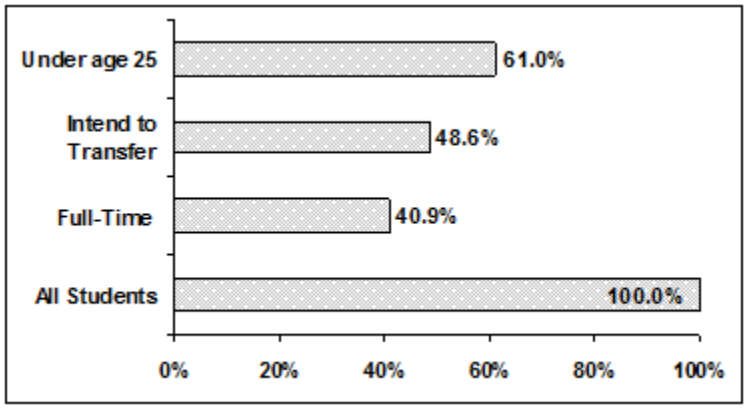


Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Characteristic/Category	Percent
Full-time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

De Anza College Spring 2010

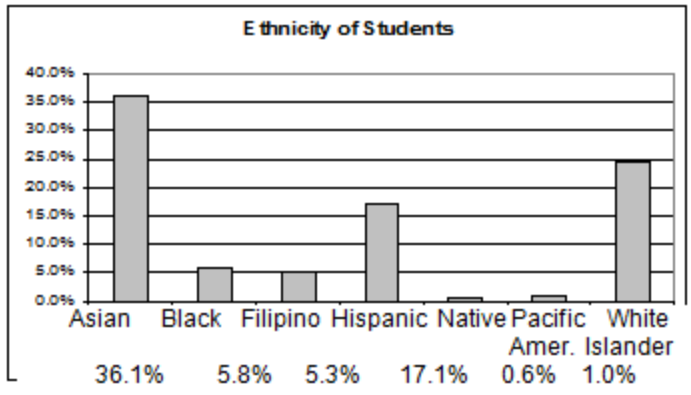


Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. Create a bar graph and not a pie chart.

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

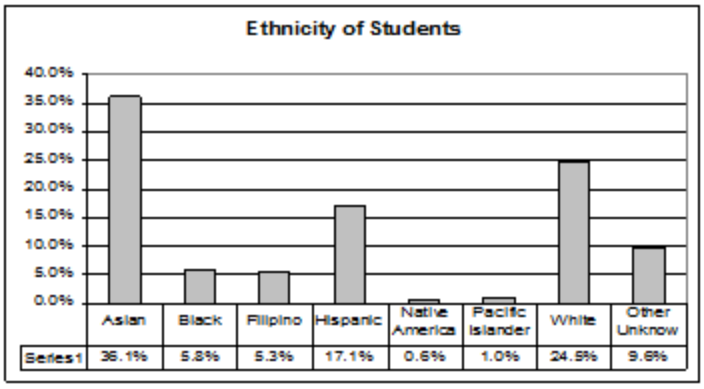
Missing Data: Ethnicity of Students De Anza College Fall Term 2007 (Census Day)



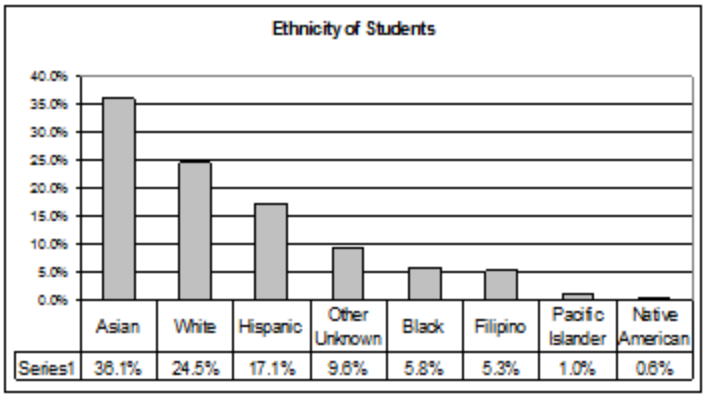
Bar graph Without Other/Unknown Category

The following graph is the same as the previous graph but the "Other/Unknown" percent (9.6%) has been added back in. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0% particularly). This is important to know when we think about what the data are telling us.

This particular bar graph can be hard to understand visually. The graph below it is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.



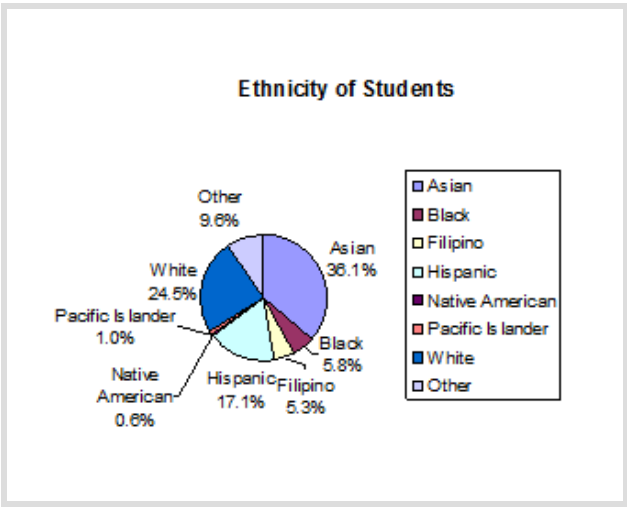
Bar Graph With Other/Unknown Category

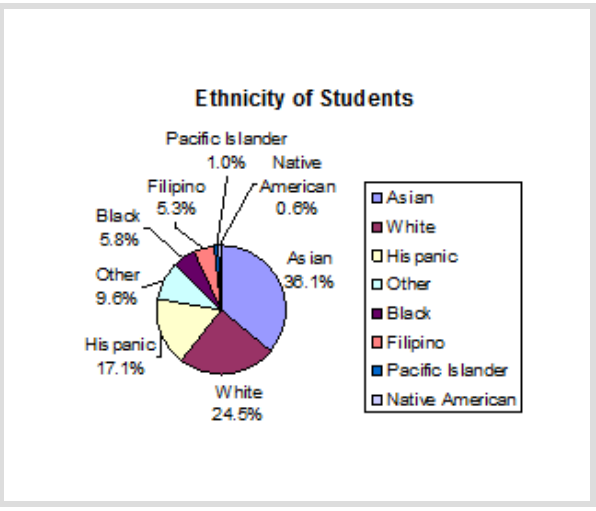


Pareto Chart With Bars Sorted By Size

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category added back in (since the percentages must add to 100%). The chart on the right is organized having the wedges by size and makes for a more visually informative graph than the unsorted, alphabetical graph on the left.





Glossary

Continuous Random Variable

A random variable (RV) whose outcomes are measured.

Example:
The height of trees in the forest is a continuous RV.

Data

A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Discrete Random Variable

A random variable (RV) whose outcomes are counted.

Qualitative Data

See **Data**.

Quantitative Data

See **Data**.

Sampling

This module introduces the concept of statistical sampling. Students are taught the difference between a simple random sample, stratified sample, cluster sample, systematic sample, and convenience sample. Example problems are provided, including an optional classroom activity.

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods.

There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of n individuals is equally likely to be chosen by any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size 3 from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out 3 names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number as shown below.

ID	Name
00	Anselmo

ID	Name
01	Bautista
02	Bayani
03	Cheng
04	Cuarismo
05	Cuningham
06	Fontecha
07	Hong
08	Hoobler
09	Jiao
10	Khan
11	King
12	Legeny
13	Lundquist
14	Macierz
15	Motogawa
16	Okimoto
17	Patel

ID	Name
18	Price
19	Quizon
20	Reyes
21	Roquero
22	Roth
23	Rowell
24	Salangsang
25	Slade
26	Stracher
27	Tallai
28	Tran
29	Wai
30	Wood

Class Roster

Lisa can either use a table of random numbers (found in many statistics books as well as mathematical handbooks) or a calculator or computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are:

.94360 .99832 .14669 .51470 .40581 .73381 .04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads .94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers .94360 and .99832 do not contain appropriate two digit numbers. However the third random number, .14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cunningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, and Cunningham, and Cuarismo.

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. For example, divide your college faculty by department. The departments are the clusters. Number each department and then choose

four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every nth piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1 - 20,000 and then use a simple random sample to pick a number that represents the first name of the sample. Then choose every 50th name thereafter until you have a total of 400 names (you might have to go back to the of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is nonrandom is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favors certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (for example, they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once using with replacement is very low.

For example, in a college population of 10,000 people, suppose you want to randomly pick a sample of 1000 for a survey. **For any particular sample of 1000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to 4 place decimals. To 4 decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement only becomes a mathematics issue when the population is small which is not that common. For example, if the population is 25 people, the sample is 10 and you are sampling **with replacement for any particular sample**,

- the chance of picking the first person is 10 out of 25 and a different second person is 9 out of 25 (you replace the first person).

If you sample **without replacement**,

- the chance of picking the first person is 10 out of 25 and then the second person (which is different) is 9 out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To 4 decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To 4 decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Example:
Exercise:

Problem:

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

1. A soccer coach selects 6 players from a group of boys aged 8 to 10, 7 players from a group of boys aged 11 to 12, and 3 players from a group of boys aged 13 to 14 to form a recreational soccer team.
2. A pollster interviews all human resource personnel in five different high tech companies.
3. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
4. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.

5. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
6. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Solution:

1. stratified
2. cluster
3. stratified
4. systematic
5. simple random
6. convenience

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will seem natural.

Example:
Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.
Suppose we take two different samples.
First, we use convenience sampling and survey 10 students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class . The amount of money they spend is as follows:
$128 $87 $173 $116 $130 $204 $147 $189 $93 $153

The second sample is taken by using a list from the P.E. department of senior citizens who take P.E. classes and taking every 5th senior citizen on the list, for a total of 10 senior citizens. They spend:
$50 $40 $36 $15 $50 $100 $40 $53 $22 $22

Exercise:

Problem:

Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Solution:

No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average part-time student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

Exercise:

Problem:

Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Solution:

No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of

part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he/she has a corresponding number. The students spend:
$180 $50 $150 $85 $260 $75 $180 $200 $200 $150

Exercise:

Problem: Is the sample biased?

Solution:

The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Optional Collaborative Classroom Exercise

Exercise:

Problem:

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

1. To find the average GPA of all students in a university, use all honor students at the university as the sample.
2. To find out the most popular cereal among young people under the age of 10, stand outside a large supermarket for three hours and speak to every 20th child under age 10 who enters the supermarket.

3. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
4. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
5. To determine the average cost of a two day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

Variation
This module discusses statistical variability within data and samples. Students will be given the opportunity to see this variability in action through participation in an optional classroom exercise. This module also has a section that discusses Critical Evaluation.

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8 16.1 15.2 14.8 15.8 15.9 16.0 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more **samples** from the same **population**, taken randomly, and having close to the same characteristics of the population are different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the

same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This **variability in samples** cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1200 to 1500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariable biased because people choose to respond or not.

Optional Collaborative Classroom Exercise

Exercise:

Problem:

Divide into groups of two, three, or four. Your instructor will give each group one 6-sided die. **Try this experiment twice.** Roll one fair die (6-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get below ("frequency" is the number of times a particular face of the die occurs):

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

First Experiment (20 rolls)

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

Second Experiment (20 rolls)

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? (Answer yes or no.) Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

Critical Evaluation

We need to critically evaluate the statistical studies we read about and analyze before accepting the results of the study. Common problems to be aware of include

- Problems with Samples: A sample should be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-Selected Samples: Responses only by people who choose to respond, such as call-in surveys are often unreliable.
- Sample Size Issues: Samples that are too small may be unreliable. Larger samples are better if possible. In some situations, small samples are unavoidable and can still be used to draw conclusions, even though larger samples are better. Examples: Crash testing cars, medical testing for rare conditions.
- Undue influence: Collecting data or asking questions in a way that influences the response.
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may both be related (correlated) because of their relationship through a different variable.
- Self-Funded or Self-Interest Studies: A study performed by a person or organization in order to support their claim. Is the study impartial?

Read the study carefully to evaluate the work. Do not automatically assume that the study is good but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.

- Misleading Use of Data: Improperly displayed graphs, incomplete data, lack of context.
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

Glossary

Population
: The collection, or set, of all individuals, objects, or measurements whose properties are being studied.

Sample
: A portion of the population understudy. A sample is representative if it characterizes the population being studied.

Answers and Rounding Off

This module briefly explains the correct way to round off answers when working with statistical data.

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round only the final answer. Do not round any intermediate results, if possible. If it becomes necessary to round intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores 4, 6, 9 is 6.3, rounded to the nearest tenth, because the data are whole numbers. Most answers will be rounded in this manner.

It is not necessary to reduce most fractions in this course. Especially in Probability Topics, the chapter on probability, it is more helpful to leave an answer as an unreduced fraction.

Frequency

This module introduces the concepts of frequency, relative frequency, and cumulative relative frequency, and the relationship between these measures. Students will have the opportunity to interpret data through the sample problems provided.

Twenty students were asked how many hours they worked per day. Their responses, in hours, are listed below:

5 6 3 3 2 4 7 5 2 3 5 6 5 4 4 3 5 2 5 3

Below is a frequency table listing the different data values in ascending order and their frequencies.

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

Frequency Table of Student Work Hours

A **frequency** is the number of times a given datum occurs in a data set. According to the table above, there are three students who work 2 hours, five students who work 3 hours, etc. The total of the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the fraction or proportion of times an answer occurs. To find the relative frequencies, divide each frequency by the total number of students in the sample - in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

Frequency Table of Student Work Hours w/ Relative Frequency

The sum of the relative frequency column is $\frac{20}{20}$, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

Frequency Table of Student Work Hours w/ Relative and Cumulative Relative Frequency

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

Note: Because of rounding, the relative frequency column may not always sum to one and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

The following table represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95 - 61.95	5	$\frac{5}{100}$ = 0.05	0.05
61.95 - 63.95	3	$\frac{3}{100}$ = 0.03	0.05 + 0.03 = 0.08
63.95 - 65.95	15	$\frac{15}{100}$ = 0.15	0.08 + 0.15 = 0.23
65.95 - 67.95	40	$\frac{40}{100}$ = 0.40	0.23 + 0.40 = 0.63
67.95 - 69.95	17	$\frac{17}{100}$ = 0.17	0.63 + 0.17 = 0.80
69.95 - 71.95	12	$\frac{12}{100}$ = 0.12	0.80 + 0.12 = 0.92
71.95 - 73.95	7	$\frac{7}{100}$ = 0.07	0.92 + 0.07 = 0.99
73.95 - 75.95	1	$\frac{1}{100}$ = 0.01	0.99 + 0.01 = 1.00
	Total = 100	Total = 1.00	

Frequency Table of Soccer Player Height

The data in this table has been **grouped** into the following intervals:

- 59.95 - 61.95 inches
- 61.95 - 63.95 inches
- 63.95 - 65.95 inches
- 65.95 - 67.95 inches
- 67.95 - 69.95 inches

- 69.95 - 71.95 inches
- 71.95 - 73.95 inches
- 73.95 - 75.95 inches

Note: This example is used again in the Descriptive Statistics chapter, where the method used to compute the intervals will be explained.

In this sample, there are **5** players whose heights are between 59.95 - 61.95 inches, **3** players whose heights fall within the interval 61.95 - 63.95 inches, **15** players whose heights fall within the interval 63.95 - 65.95 inches, **40** players whose heights fall within the interval 65.95 - 67.95 inches, **17** players whose heights fall within the interval 67.95 - 69.95 inches, **12** players whose heights fall within the interval 69.95 - 71.95, 7 players whose height falls within the interval 71.95 - 73.95, and **1** player whose height falls within the interval 73.95 - 75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Example:
Exercise:

Problem:

From the table, find the percentage of heights that are less than 65.95 inches.

Solution:

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are 5 + 3 + 15 = 23 males whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23%. This percentage is the cumulative relative frequency entry in the third row.

Example:

Exercise:

Problem:

From the table, find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution:

Add the relative frequencies in the second and third rows: 0.03 + 0.15 = 0.18 or 18%.

Example:
Exercise:

Problem:

Use the table of heights of the 100 male semiprofessional soccer players. Fill in the blanks and check your answers.

1. The percentage of heights that are from 67.95 to 71.95 inches is:
2. The percentage of heights that are from 67.95 to 73.95 inches is:
3. The percentage of heights that are more than 65.95 inches is:
4. The number of players in the sample who are between 61.95 and 71.95 inches tall is:
5. What kind of data are the heights?
6. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Solution:

1. 29%
2. 36%
3. 77%

4. 87
5. quantitative continuous
6. get rosters from each team and choose a simple random sample from each

Optional Collaborative Classroom Exercise

Exercise:

Problem:

In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

1. What percentage of the students in your class has 0 siblings?
2. What percentage of the students has from 1 to 3 siblings?
3. What percentage of the students has fewer than 3 siblings?

Example:
Nineteen people were asked how many miles, to the nearest mile they commute to work each day. The data are as follows:
2 5 7 3 2 10 18 15 20 7 10 18 5 12 13 12 4 5 10
The following table was produced:

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{4}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

Frequency of Commuting Distances

Exercise:

Problem:

1. Is the table correct? If it is not correct, what is wrong?
2. True or False: Three percent of the people surveyed commute 3 miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
3. What fraction of the people surveyed commute 5 or 7 miles?
4. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between 5 and 13 miles (does not include 5 and 13 miles)?

Solution:

1. No. Frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
2. False. Frequency for 3 miles should be 1; for 2 miles (left out), 2. Cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.
3. $\frac{5}{19}$
4. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

Glossary

Frequency
: The number of times a value of the data occurs.

Relative Frequency
: The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes.

Cumulative Relative Frequency
: The term applies to an ordered set of observations from smallest to largest. The Cumulative Relative Frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Summary
This module provides an outline/review of key concepts related to statistical sampling and data.

Statistics

- Deals with the collection, analysis, interpretation, and presentation of data

Probability

- Mathematical tool used to study randomness

Key Terms

- Population
- Parameter
- Sample
- Statistic
- Variable
- Data

Types of Data

- Quantitative Data (a number)
    - Discrete (You count it.)
    - Continuous (You measure it.)
- Qualitative Data (a category, words)

Sampling

- **With Replacement**: A member of the population may be chosen more than once
- **Without Replacement**: A member of the population may be chosen only once

Random Sampling

- Each member of the population has an equal chance of being selected

Sampling Methods

- Random
    - Simple random sample
    - Stratified sample
    - Cluster sample
    - Systematic sample
- Not Random
    - Convenience sample

Frequency (freq. or f)

- The number of times an answer occurs

Relative Frequency (rel. freq. or RF)

- The proportion of times an answer occurs
- Can be interpreted as a fraction, decimal, or percent

Cumulative Relative Frequencies (cum. rel. freq. or cum RF)

- An accumulation of the previous relative frequencies

Practice: Sampling and Data
This module provides an opportunity for students to practice concepts related to statistical sampling and data. Given a sample data set, the student will practice constructing frequency tables, differentiating between key terms, and comparing sampling techniques.

Student Learning Outcomes

- The student will construct frequency tables.
- The student will differentiate between key terms.
- The student will compare sampling techniques.

Given

Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average(mean) length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A3 4 11 15 16 17 22 44 37 16 14 24 25 15 26 27 33 29 35 44 13 21 22 10 12 8 40 32 26 27 31 34 29 17 8 24 18 47 33 34

Researcher B3 14 11 5 16 17 28 41 31 18 14 14 26 25 21 22 31 2 35 44 23 21 21 16 12 18 41 22 16 25 33 34 29 13 18 24 23 42 33 29

Organize the Data

Complete the tables below using the data provided.

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5 - 6.5			
6.5 - 12.5			
12.5 - 18.5			
18.5 - 24.5			
24.5 - 30.5			
30.5 - 36.5			
36.5 - 42.5			
42.5 - 48.5			

Researcher A

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5 - 6.5			
6.5 - 12.5			
12.5 - 18.5			

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
18.5 - 24.5			
24.5 - 30.5			
30.5 - 36.5			
36.5 - 42.5			
42.5 - 48.5			

Researcher B

Key Terms

Define the key terms based upon the above example for Researcher A.
Exercise:

Problem: Population

Exercise:

Problem: Sample

Exercise:

Problem: Parameter

Exercise:

Problem: Statistic

Exercise:

Problem: Variable

Exercise:

Problem: Data

Discussion Questions

Discuss the following questions and then answer in complete sentences.

Exercise:

Problem: List two reasons why the data may differ.

Exercise:

Problem:

Can you tell if one researcher is correct and the other one is incorrect? Why?

Exercise:

Problem: Would you expect the data to be identical? Why or why not?

Exercise:

Problem: How could the researchers gather random data?

Exercise:

Problem:

Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?

Exercise:

Problem:

Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?

Homework
This module presents students with a number of problems related to statistical sampling and data. In particular, students are asked to demonstrate understanding of concepts such as frequency, relative frequency, and cumulative relative frequency, random samples, quantitative vs. qualitative data, continuous vs. discrete data, and other key terms related to sampling and data.

Exercise:

Problem: For each item below:

- **i**Identify the type of data (quantitative - discrete, quantitative - continuous, or qualitative) that would be used to describe a response.
- **ii**Give an example of the data.

- **a**Number of tickets sold to a concert
- **b**Amount of body fat
- **c**Favorite baseball team
- **d**Time in line to buy groceries
- **e**Number of students enrolled at Evergreen Valley College
- **f**Most–watched television show
- **g**Brand of toothpaste
- **h**Distance to the closest movie theatre
- **i**Age of executives in Fortune 500 companies
- **j**Number of competing computer spreadsheet software packages

---

Solution:

- **a**quantitative - discrete
- **b**quantitative - continuous
- **c**qualitative
- **d**quantitative - continuous
- **e**quantitative - discrete
- **f**qualitative
- **g**qualitative

- **h**quantitative - continuous
- **i**quantitative - continuous
- **j**quantitative - discrete

Exercise:

Problem:

Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		
3			

Part-time Student Course Loads

- **a**Fill in the blanks in the table above.
- **b**What percent of students take exactly two courses?
- **c**What percent of students take one or two courses?

Exercise:

Problem:

Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnoses. The (incomplete) results are shown below:

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Freq.
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

Flossing Frequency for Adults with Gum Disease

- **a**Fill in the blanks in the table above.
- **b**What percent of adults flossed six times per week?
- **c**What percent flossed at most three times per week?

Solution:

- **a** Cum. Rel. Freq. for 0 is 0.4500
  Rel. Freq. for 1 is 0.3000 and Cum. Rel. Freq. for 1 or less is 0.7500

Freq. for 3 is 11 and Rel. Freq. is 0.1833
Cum. Rel. Freq. for 6 or less is 0.9833
Cum. Rel. Freq. for 7 or less is 1

- **b**5.00%
- **c**93.33%

Exercise:

Problem:

A fitness center is interested in the mean amount of time a client exercises in the center each week. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Exercise:

Problem:

Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to optimally plan their ski classes. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Solution:

- **a** Children who take ski or snowboard lessons
- **b** A group of these children
- **c** The population mean
- **d** The sample mean
- **e** X = the age of one child who takes the first ski or snowboard lesson
- **f** Values for X, such as 3, 7, etc.

Exercise:

Problem:

A cardiologist is interested in the mean recovery period for her patients who have had heart attacks. Define the following in terms of the study. Give examples where appropriate.

- **a** Population
- **b** Sample
- **c** Parameter
- **d** Statistic
- **e** Variable
- **f** Data

Exercise:

Problem:

Insurance companies are interested in the mean health costs each year for their clients, so that they can determine the costs of health insurance. Define the following in terms of the study. Give examples where appropriate.

- **a** Population
- **b** Sample
- **c** Parameter
- **d** Statistic

- **e**Variable
- **f**Data

Solution:

- **a**The clients of the insurance companies
- **b**A group of the clients
- **c**The mean health costs of the clients
- **d**The mean health costs of the sample
- **e**X = the health costs of one client
- **f**Values for X, such as 34, 9, 82, etc.

Exercise:

Problem:

A politician is interested in the proportion of voters in his district that think he is doing a good job. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Exercise:

Problem:

A marriage counselor is interested in the proportion the clients she counsels that stay married. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample

- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Solution:

- **a**All the clients of the counselor
- **b**A group of the clients
- **c**The proportion of all her clients who stay married
- **d**The proportion of the sample who stay married
- **e**X = the number of couples who stay married
- **f**yes, no

Exercise:

Problem:

Political pollsters may be interested in the proportion of people that will vote for a particular cause. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Exercise:

Problem:

A marketing company is interested in the proportion of people that will buy a particular product. Define the following in terms of the study. Give examples where appropriate.

- **a** Population
- **b** Sample
- **c** Parameter
- **d** Statistic
- **e** Variable
- **f** Data

Solution:

- **a** All people (maybe in a certain geographic area, such as the United States)
- **b** A group of the people
- **c** The proportion of all people who will buy the product
- **d** The proportion of the sample who will buy the product
- **e** X = the number of people who will buy it
- **f** buy, not buy

Exercise:

Problem:

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys 6 flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

- **a** Using complete sentences, list three things wrong with the way the survey was conducted.
- **b** Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

Exercise:

Problem:

Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in 3 – 5 complete sentences. Make the description detailed.

Exercise:

Problem:

Suppose you want to determine the mean number of cans of soda drunk each month by persons in their twenties. Describe a possible sampling method in 3 - 5 complete sentences. Make the description detailed.

Exercise:

Problem:

771 distance learning students at Long Beach City College responded to surveys in the 2010-11 academic year. Highlights of the summary report are listed in the table below. (Source: http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus).

Have computer at home	96%
Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%
Live at least 16 miles from campus	13%

Took DL courses to fulfill transfer requirements	71%

LBCC Distance Learning Survey Results

- **a**What percent of the students surveyed do not have a computer at home?
- **b**About how many students in the survey live at least 16 miles from campus?
- **c**If the same survey was done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

Solution:

- **a**4%
- **b**100

Exercise:

Problem:

Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows:

2 5 7 2 2 10 20 15 0 7 0 20 5 12 15 12 4 5 10

The following table was produced:

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
0	2	$\frac{2}{19}$	0.1053

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
2	3	$\frac{3}{19}$	0.2632
4	1	$\frac{1}{19}$	0.3158
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.5789
10	2	$\frac{2}{19}$	0.6842
12	2	$\frac{2}{19}$	0.7895
15	1	$\frac{1}{19}$	0.8421
20	1	$\frac{1}{19}$	1.0000

Frequency of Immigrant Survey Responses

- **a**Fix the errors on the table. Also, explain how someone might have arrived at the incorrect number(s).
- **b**Explain what is wrong with this statement: “47 percent of the people surveyed have lived in the U.S. for 5 years.”
- **c**Fix the statement above to make it correct.
- **d**What fraction of the people surveyed have lived in the U.S. 5 or 7 years?
- **e**What fraction of the people surveyed have lived in the U.S. at most 12 years?
- **f**What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
- **g**What fraction of the people surveyed have lived in the U.S. from 5 to 20 years, inclusive?

Exercise:

Problem:

A “random survey” was conducted of 3274 people of the “microprocessor generation” (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had $2000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users. (*Source: San Jose Mercury News*)

- **a**Do you consider the sample size large enough for a study of this type? Why or why not?
- **b**Based on your “gut feeling,” do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?

Additional information: The survey was reported by Intel Corporation of individuals who visited the Los Angeles Convention Center to see the Smithsonian Institure's road show called “America’s Smithsonian.”

- **c**With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
- **d**With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

Exercise:

Problem:

- **a**List some practical difficulties involved in getting accurate results from a telephone survey.
- **b**List some practical difficulties involved in getting accurate results from a mailed survey.
- **c**With your classmates, brainstorm some ways to overcome these problems if you needed to conduct a phone or mail survey.

Try these multiple choice questions

The next four questions refer to the following: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

Exercise:

Problem: What is the population she is interested in?

- **A** All Lake Tahoe Community College students
- **B** All Lake Tahoe Community College English students
- **C** All Lake Tahoe Community College students in her classes
- **D** All Lake Tahoe Community College math students

Solution:

D

Exercise:

Problem: Consider the following:

X = **number of days a Lake Tahoe Community College math student is absent**

In this case, X is an example of a:

- **A** Variable
- **B** Population
- **C** Statistic
- **D** Data

Solution:

A

Exercise:

Problem:

The instructor takes her sample by gathering data on 5 randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

- **A** Cluster sampling
- **B** Stratified sampling
- **C** Simple random sampling
- **D** Convenience sampling

Solution:

B

Exercise:

Problem:

The instructor’s sample produces an mean number of days absent of 3.5 days. This value is an example of a

- **A** Parameter
- **B** Data
- **C** Statistic
- **D** Variable

Solution:

C

The next two questions refer to the following relative frequency table on hurricanes that have made direct hits on the U.S between 1851 and 2004. Hurricanes are given a strength category rating based on the minimum wind speed generated by the storm. (http://www.nhc.noaa.gov/gifs/table5.gif)

Category	Number of Direct Hits	Relative Frequency	Cumulative Frequency
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

Frequency of Hurricane Direct Hits

Exercise:

Problem:

What is the relative frequency of direct hits that were category 4 hurricanes?

- **A**0.0768
- **B**0.0659
- **C**0.2601
- **D**Not enough information to calculate

Solution:

B

Exercise:

Problem:

What is the relative frequency of direct hits that were AT MOST a category 3 storm?

- **A**0.3480
- **B**0.9231
- **C**0.2601
- **D**0.3370

Solution:

B

The next three questions refer to the following: A study was done to determine the age, number of times per week and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

Exercise:

Problem: "‘Number of times per week’" is what type of data?

- **A**qualitative
- **B**quantitative - discrete
- **C**quantitative - continuous

Solution:

B

Exercise:

Problem: The sampling method was:

- **A**simple random
- **B**systematic
- **C**stratified
- **D**cluster

Solution:

B

Exercise:

Problem: "‘Duration (amount of time)’" is what type of data?

- **A**qualitative
- **B**quantitative - discrete
- **C**quantitative - continuous

Solution:

C

Exercises 28 and 29 are not multiple choice exercises.

Exercise:

Problem:

Name the sampling method used in each of the following situations:

- **A**A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport’s service. She does not ask travelers who are hurrying through the airport with their

hands full of luggage, but instead asks all travelers sitting near gates and who are not taking naps while they wait.

- **B**A teacher wants to know if her students are doing homework so she randomly selects rows 2 and 5, and then calls on all students in row 2 and all students in row 5 to present the solution to homework problems to the class.
- **C**The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out which asks for information about age, as well as about other variables of interest.
- **D**The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether the books are checked out by an adult or a child. She records this data for every 4th patron who checks out books.
- **E**A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked who he/she intends to vote for and whether the debate changed his/her opinion of the candidates.

** Contributed by Roberta Bloom

Solution:

- **A**Convenience
- **B**Cluster
- **C**Stratified
- **D**Systematic
- **E**Simple Random

Exercise:

Problem:

Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the college term. If the book is not available, then a student would not be able to get the textbook at all, or might get a delayed delivery if the book is back ordered.

A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook for each of the following 7 subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these 7 textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers.

Write an analysis of his study that addresses the following issues: Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

** Contributed by Roberta Bloom

Solution:

The answer below contains some of the issues that students might discuss for this problem. Individual student's answers may also identify other issues that pertain to this problem that are not included in the answer below.

The sample is not representative of the population of all college textbooks. Two reasons why it is not representative are that he only sampled 7 subjects and he only investigated one textbook in each subject. There are several possible sources of bias in the study. The 7 subjects that he investigated are all in mathematics and the sciences; there are many subjects in the humanities, social sciences, and many other subject areas, (for example: literature, art, history, psychology, sociology, business) that he did not investigate at all. It may be that different subject areas exhibit different patterns of textbook availability, but his sample would not detect such results.

He also only looked at the most popular textbook in each of the subjects he investigated. The availability of the most popular textbooks may differ from the availability of other textbooks in one of two ways:

- the most popular textbooks may be more readily available online, because more new copies are printed and more students nationwide selling back their used copies OR
- the most popular textbooks may be harder to find available online, because more student demand exhausts the supply more quickly.

In reality, many college students do not use the most popular textbook in their subject, and this study gives no useful information about the situation for those less popular textbooks.

He could improve this study by

- expanding the selection of subjects he investigates so that it is more representative of all subjects studied by college students and
- expanding the selection of textbooks he investigates within each subject to include a mixed representation of both the popular and less popular textbooks.

Lab 1: Data Collection
This lab allows students to practice and demonstrate techniques used to generate systematic samples. Students will have the opportunity to create relative frequency tables and interpret results based on different data groupings.

Class Time:

Names:

Student Learning Outcomes

- The student will demonstrate the systematic sampling technique.
- The student will construct Relative Frequency Tables.
- The student will interpret results and their differences from different data groupings.

Movie Survey

Ask five classmates from a different class how many movies they saw last month at the theater. Do not include rented movies.

1. Record the data
2. In class, randomly pick one person. On the class list, mark that person's name. Move down four people's names on the class list. Mark that person's name. Continue doing this until you have marked 12 people's names. You may need to go back to the start of the list. For each marked name record below the five data values. You now have a total of 60 data values.
3. For each name marked, record the data:

_____	_____	_____	_____	_____	_____	_____	_____	____
_____	_____	_____	_____	_____	_____	_____	_____	____
_____	_____	_____	_____	_____	_____	_____	_____	____
_____	_____	_____	_____	_____	_____	_____	_____	____
_____	_____	_____	_____	_____	_____	_____	_____	____
_____	_____	_____	_____	_____	_____	_____	_____	____

Order the Data

Complete the two relative frequency tables below using your class data.

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0			

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
1			
2			
3			
4			
5			
6			
7+			

Frequency of Number of Movies Viewed

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0-1			
2-3			
4-5			
6-7+			

Frequency of Number of Movies Viewed

1. Using the tables, find the percent of data that is at most 2. Which table did you use and why?
2. Using the tables, find the percent of data that is at most 3. Which table did you use and why?
3. Using the tables, find the percent of data that is more than 2. Which table did you use and why?
4. Using the tables, find the percent of data that is more than 3. Which table did you use and why?

Discussion Questions

1. Is one of the tables above "more correct" than the other? Why or why not?
2. In general, why would someone group the data in different ways? Are there any advantages to either way of grouping the data?
3. Why did you switch between tables, if you did, when answering the question above?

Lab 2: Sampling Experiment
This module provides students an opportunity to practice data sampling techniques. Students will demonstrate the simple random, systematic, stratified, and cluster sampling techniques, and explain the details of each procedure. This lab is based on one originally created by Carol Olmstead.

Class Time:

Names:

Student Learning Outcomes

- The student will demonstrate the simple random, systematic, stratified, and cluster sampling techniques.
- The student will explain each of the details of each procedure used.

In this lab, you will be asked to pick several random samples. In each case, describe your procedure briefly, including how you might have used the random number generator, and then list the restaurants in the sample you obtained

Note: The following section contains restaurants stratified by city into columns and grouped horizontally by entree cost (clusters).

A Simple Random Sample

Pick a **simple random sample** of 15 restaurants.

1. Describe the procedure:

2.

1. _____	6. _____	11. _____
2. _____	7. _____	12. _____
3. _____	8. _____	13. _____
4. _____	9. _____	14. _____

5. _____	10. _____	15. _____

A Systematic Sample

Pick a **systematic sample** of 15 restaurants.

1. Describe the procedure:

2.

1. _____	6. _____	11. _____
2. _____	7. _____	12. _____
3. _____	8. _____	13. _____
4. _____	9. _____	14. _____
5. _____	10. _____	15. _____

A Stratified Sample

Pick a **stratified sample**, by city, of 20 restaurants. Use 25% of the restaurants from each stratum. Round to the nearest whole number.

1. Describe the procedure:

2.

1. _____	6. _____	11. _____	16. _____
2. _____	7. _____	12. _____	17. _____
3. _____	8. _____	13. _____	18. _____
4. _____	9. _____	14. _____	19. _____
5. _____	10. _____	15. _____	20. _____

A Stratified Sample

Pick a **stratified sample**, by entree cost, of 21 restaurants. Use 25% of the restaurants from each stratum. Round to the nearest whole number.

1. Describe the procedure:

2.

1. _____	6. _____	11. _____	16. _____
2. _____	7. _____	12. _____	17. _____
3. _____	8. _____	13. _____	18. _____
4. _____	9. _____	14. _____	19. _____
5. _____	10. _____	15. _____	20. _____
			21. _____

A Cluster Sample

Pick a **cluster sample** of restaurants from two cities. The number of restaurants will vary.

1. Describe the procedure:

2.

1. _____	6. _____	11. _____	16. _____	21. _____
2. _____	7. _____	12. _____	17. _____	22. _____
3. _____	8. _____	13. _____	18. _____	23. _____
4.	9.	14.	19.	24.

_____	_____	_____	_____	_____
5. _____	10. _____	15. _____	20. _____	25. _____

Restaurants Stratified by City and Entree Cost

Entree Cost →	**Under $10**	**$10 to under $15**	**$15 to under $20**	**Over $20**
San Jose	El Abuelo Taq, Pasta Mia, Emma's Express, Bamboo Hut	Emperor's Guard, Creekside Inn	Agenda, Gervais, Miro's	Blake's, Eulipia, Hayes Mansion, Germania
Palo Alto	Senor Taco, Olive Garden, Taxi's	Ming's, P.A. Joe's, Stickney's	Scott's Seafood, Poolside Grill, Fish Market	Sundance Mine, Maddalena's, Spago's
Los Gatos	Mary's Patio, Mount Everest, Sweet Pea's, Andele Taqueria	Lindsey's, Willow Street	Toll House	Charter House, La Maison Du Cafe
Mountain View	Maharaja, New Ma's, Thai-Rific, Garden Fresh	Amber Indian, La Fiesta, Fiesta del Mar, Dawit	Austin's, Shiva's, Mazeh	Le Petit Bistro

Entree Cost →	**Under $10**	**$10 to under $15**	**$15 to under $20**	**Over $20**
Cupertino	Hobees, Hung Fu, Samrat, Panda Express	Santa Barb. Grill, Mand. Gourmet, Bombay Oven, Kathmandu West	Fontana's, Blue Pheasant	Hamasushi, Helios
Sunnyvale	Chekijababi, Taj India, Full Throttle, Tia Juana, Lemon Grass	Pacific Fresh, Charley Brown's, Cafe Cameroon, Faz, Aruba's	Lion & Compass, The Palace, Beau Sejour	
Santa Clara	Rangoli, Armadillo Willy's, Thai Pepper, Pasand	Arthur's, Katie's Cafe, Pedro's, La Galleria	Birk's, Truya Sushi, Valley Plaza	Lakeside, Mariani's

Restaurants Used in Sample

Note:*The original lab was designed and contributed by Carol Olmstead.*

Descriptive Statistics

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: stemplots, histograms and boxplots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Introduction

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **"Descriptive Statistics"**. You will learn to calculate, and even more importantly, to interpret these measurements and graphs.

Displaying Data

This module provides a brief introduction into the ways graphs and charts can be used to provide visual representations of data.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample. The graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly.

Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar chart, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), pie charts, and the boxplot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs and bar graphs. Our emphasis will be on histograms and boxplots.

Stem and Leaf Graphs (Stemplots), Line Graphs and Bar Graphs
This module introduces the use of stem-and-leaf graphs (stemplots), line graphs and bar graphs for describing a set of data visually.

One simple graph, the **stem-and-leaf graph** or **stem plot**, comes from the field of exploratory data analysis.It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a **final significant digit**. For example, 23 has stem 2 and leaf 3. Four hundred thirty-two (432) has stem 43 and leaf 2. Five thousand four hundred thirty-two (5,432) has stem 543 and leaf 2. The decimal 9.3 has stem 9 and leaf 3. Write the stems in a vertical line from smallest the largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example:
For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):
33424949535555616367686869697273747880838888889092949494949610 0

Stem	Leaf
3	3
4	299
5	355
6	1378899

Stem	Leaf
7	2348
8	03888
9	0244446
10	0

Stem-and-Leaf Diagram

The stem plot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% of the scores were in the 90's or 100, a fairly high number of As.

The stem plot is a quick way to graph and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers. In the example above, there were no outliers.

Example:
Create a stem plot using the data:
1.11.52.32.52.73.23.33.33.53.84.0 4.24.54.54.74.85.55.66.56.712.3
The data are the distance (in kilometers) from a home to the nearest supermarket.

Exercise:

Problem:

1. Are there any values that might possibly be outliers?
2. Do the data seem to have any concentration of values?

Note: The leaves are to the right of the decimal.

Solution:

The value 12.3 may be an outlier. Values appear to concentrate at 3 and 4 kilometers.

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	

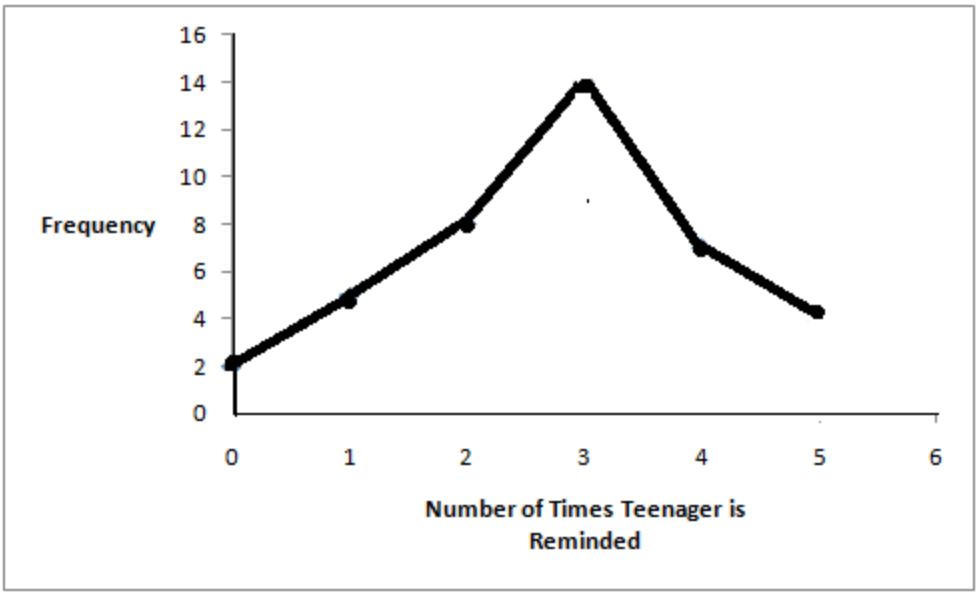
Stem	Leaf
9	
10	
11	
12	3

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in the example, the **x-axis** consists of **data values** and the **y-axis** consists of **frequency points**. The frequency points are connected.

Example:
In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his/her chores. The results are shown in the table and the line graph.

Number of times teenager is reminded	Frequency
0	2
1	5

Number of times teenager is reminded	Frequency
2	8
3	14
4	7
5	4

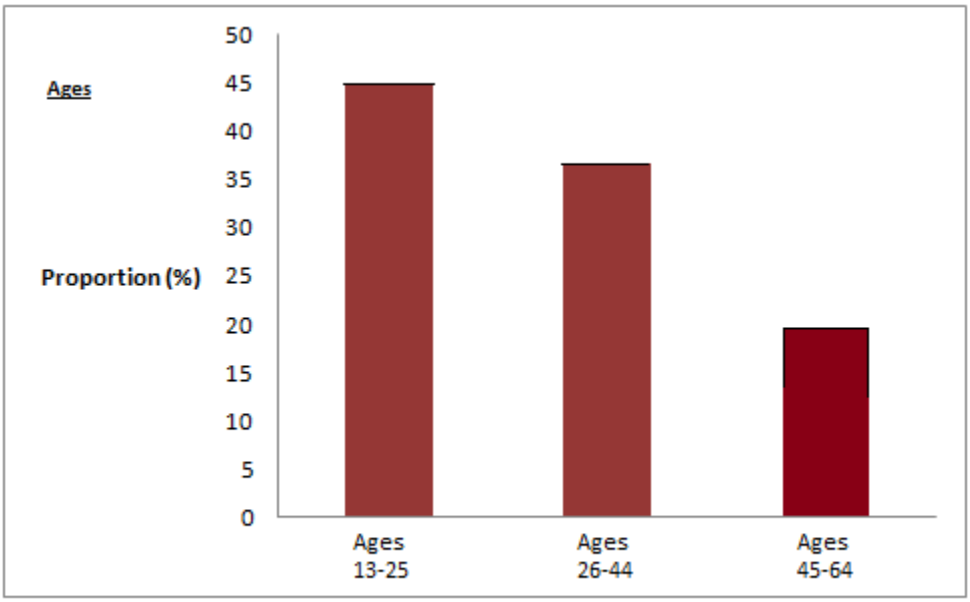


Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes and they can be vertical or horizontal.

The **bar graph** shown in **Example 4** has age groups represented on the **x-axis** and proportions on the **y-axis**.

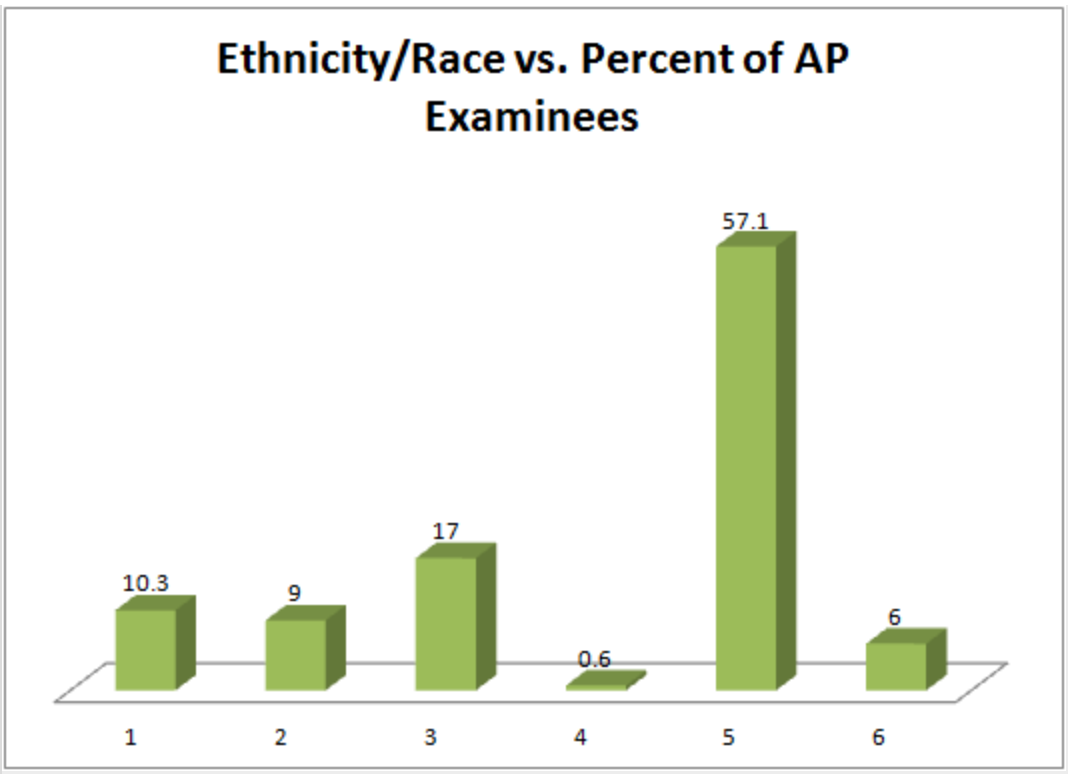
Example:
By the end of 2011, in the United States, Facebook had over 146 million users. The table shows three age groups, the number of users in each age group and the proportion (%) of users in each age group. **Source:** ***http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/***

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13 - 25	65,082,280	45%
26 - 44	53,300,200	36%
45 - 64	27,885,100	19%



Example:
The columns in the table below contain the race/ethnicity of U.S. Public Schools: High School Class of 2011, percentages for the Advanced Placement Examinee Population for that class and percentages for the Overall Student Population. The 3-dimensional graph shows the Race/Ethnicity of U.S. Public Schools (qualitative data) on the **x-axis** and Advanced Placement Examinee Population percentages on the **y-axis**. (**Source: http://www.collegeboard.com** and **Source: http://apreport.collegeboard.org/goals-and-findings/promoting-equity**)

Race/Ethnicity	**AP Examinee Population**	**Overall Student Population**
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%



Go to Outcomes of Education Figure 22 for an example of a bar graph that shows unemployment rates of persons 25 years and older for 2009.

Note: This book contains instructions for constructing a **histogram** and a **box plot** for the TI-83+ and TI-84 calculators. You can find additional instructions for using these calculators on the Texas Instruments (TI) website.

Glossary

Outlier
: An observation that does not fit the rest of the data.

Histograms

This module provides an overview of Descriptive Statistics: Histogram as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A **histogram** consists of contiguous boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either **Frequency** or **relative frequency**. The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data. (The next section tells you how to calculate the center and the spread.)

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (In the chapter on Sampling and Data, we defined frequency as the number of times an answer occurs.) If:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then:

Equation:

$$\text{RF} = \frac{f}{n}$$

For example, if 3 students in Mr. Ahab's English class of 40 students received from 90% to 100%, then,

$f = 3$, $n = 40$, and $\text{RF} = \frac{f}{n} = \frac{3}{40} = 0.075$

Seven and a half percent of the students received 90% to 100%. Ninety percent to 100 % are quantitative measures.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of from 5 to 15 bars or classes for clarity. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 - 0.05 = 6.05). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 (1.5 - 0.005 = 1.495). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 (1.0 - .0005 = 0.9995). If all the data happen to be integers and the smallest value is 2, then a convenient starting point is 1.5 (2 - 0.5 = 1.5). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary.

Example:
The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data since height is measured.
60 60.5 61 61 61.5
63.5 63.5 63.5
64 64 64 64 64 64 64 64.5 64.5 64.5 64.5 64.5 64.5 64.5 64.5
66 66 66 66 66 66 66 66 66 66 66 66.5 66.5 66.5 66.5 66.5 66.5 66.5 66.5
66.5 66.5 66.5 67 67 67 67 67 67 67 67 67 67 67 67.5 67.5 67.5 67.5
67.5 67.5 67.5
68 68 69 69 69 69 69 69 69 69 69 69 69.5 69.5 69.5 69.5 69.5
70 70 70 70 70 70 70.5 70.5 70.5 71 71 71
72 72 72 72.5 72.5 73 73.5
74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.
60 - 0.05 = 59.95 which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.
The largest value is 74. 74+ 0.05 = 74.05 is the ending value.
Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose 8 bars.
Equation:

$$\frac{74.05 - 59.95}{8} = 1.76$$

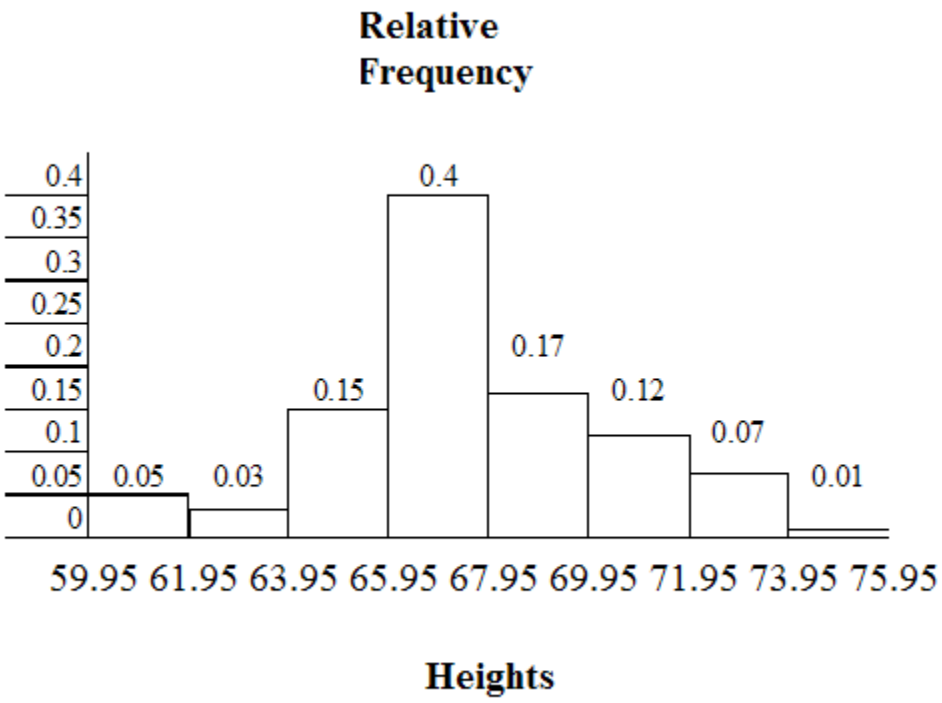
Note: We will round up to 2 and make each bar or class interval 2 units wide. Rounding up to 2 is one way to prevent a value from falling on a boundary. Rounding to the next number is necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work.

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95

- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95 - 61.95. The heights that are 63.5 are in the interval 61.95 - 63.95. The heights that are 64 through 64.5 are in the interval 63.95 - 65.95. The heights 66 through 67.5 are in the interval 65.95 - 67.95. The heights 68 through 69.5 are in the interval 67.95 - 69.95. The heights 70 through 71 are in the interval 69.95 - 71.95. The heights 72 through 73.5 are in the interval 71.95 - 73.95. The height 74 is in the interval 73.95 - 75.95.
The following histogram displays the heights on the x-axis and relative frequency on the y-axis.



Example:
The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is discrete data since books

are counted.
1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4
5 5 5 5 5
6 6
Eleven students buy 1 book. Ten students buy 2 books. Sixteen students buy 3 books. Six students buy 4 books. Five students buy 5 books. Two students buy 6 books.
Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Exercise:

Problem:

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6 and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _____ to _____, the 5 in the middle of the interval from _____ to _____, and the _____ in the middle of the interval from _____ to _____ .

Solution:

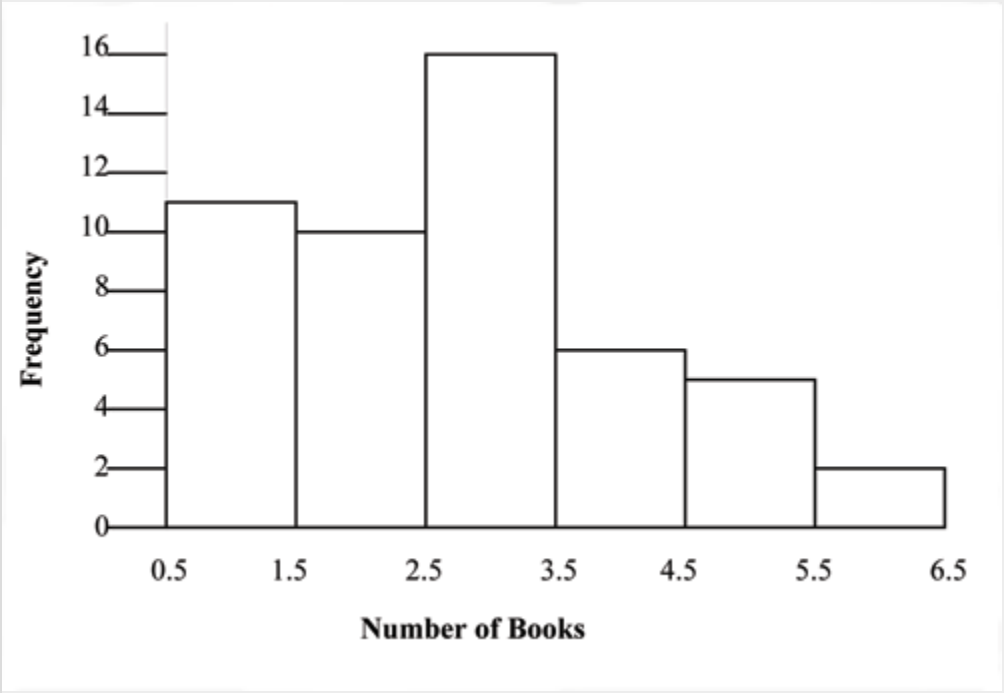
- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

Calculate the number of bars as follows:

Equation:

$$\frac{6.5 - 0.5}{\text{bars}} = 1$$

where 1 is the width of a bar. Therefore, $\text{bars} = 6$.
The following histogram displays the number of books on the x-axis and the frequency on the y-axis.



Using the TI-83, 83+, 84, 84+ Calculator Instructions
Go to the Appendix (14:Appendix) in the menu on the left. There are calculator instructions for entering data and for creating a customized histogram. Create the histogram for Example 2.

- Press Y=. Press CLEAR to clear out any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6
- Into L2, enter 11, 10, 16, 6, 5, 2
- Press WINDOW. Make Xmin = .5, Xmax = 6.5, Xscl = (6.5 - .5)/6, Ymin = -1, Ymax = 20, Yscl = 1, Xres = 1

- Press 2nd Y=. Start by pressing 4:Plotsoff ENTER.
- Press 2nd Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rd picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2).
- Press GRAPH
- Use the TRACE key and the arrow keys to examine the histogram.

Optional Collaborative Exercise

Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals. Discuss, also, the shape of the histogram.

Record the data, in dollars (for example, 1.25 dollars).

Construct a histogram.

Glossary

Frequency
: The number of times a value of the data occurs.

Relative Frequency
: The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes.

Box Plots

Box plots or **box-whisker plots** give a good graphical image of the concentration of the data. They also show how far from most of the data the extreme values are. The box plot is constructed from five values: the smallest value, the first quartile, the median, the third quartile, and the largest value. The median, the first quartile, and the third quartile will be discussed here, and then again in the section on measuring data in this chapter. We use these values to compare how close other data values are to them.

The **median**, a number, is a way of measuring the "center" of the data. You can think of the median as the "middle value," although it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger. For example, consider the following data:

1 11.5 6 7.2 4 8 9 10 6.8 8.3 2 2 10 1

Ordered from smallest to largest:

1 1 2 2 4 6 **6.8 7.2** 8 8.3 9 10 10 11.5

The median is between the 7th value, 6.8, and the 8th value 7.2. To find the median, add the two values together and divide by 2.
Equation:

$$\frac{6.8 + 7.2}{2} = 7$$

The median is 7. Half of the values are smaller than 7 and half of the values are larger than 7.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile is the middle value of the lower half of

the data and the third quartile is the middle value of the upper half of the data. To get the idea, consider the same data set shown above:

1 1 2 2 4 6 6.8 7.2 8 8.3 9 10 10 11.5

The median or **second quartile** is 7. The lower half of the data is 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is 2.

1 1 2 **2** 4 6 6.8

The number 2, which is part of the data, is the **first quartile**. One-fourth of the values are the same or less than 2 and three-fourths of the values are more than 2.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is 9.

7.2 8 8.3 **9** 10 10 11.5

The number 9, which is part of the data, is the **third quartile**. Three-fourths of the values are less than 9 and one-fourth of the values are more than 9.

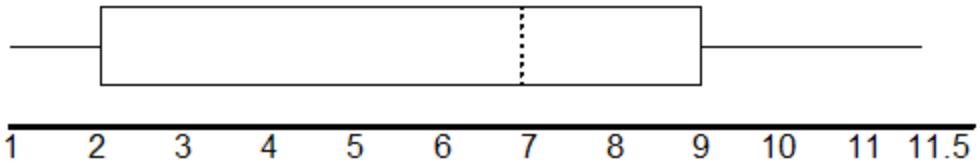
To construct a box plot, use a horizontal number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. **The middle fifty percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values. The box plot gives a good quick picture of the data.

Note: You may encounter box and whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider the following data:

1 1 2 2 4 6 6.8 7.2 8 8.3 9 10 10 11.5

The first quartile is 2, the median is 7, and the third quartile is 9. The smallest value is 1 and the largest value is 11.5. The box plot is constructed as follows (see calculator instructions in the back of this book or on the TI web site):



The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

Example:
The following data are the heights of 40 students in a statistics class.
59 60 61 62 62 63 63 64 64 64 65 65 65 65 65 65 65 65 66 66 67 67 68 68 69 70 70 70 70 70 71 71 72 72 73 74 74 75 77
Construct a box plot:
Using the TI-83, 83+, 84, 84+ Calculator

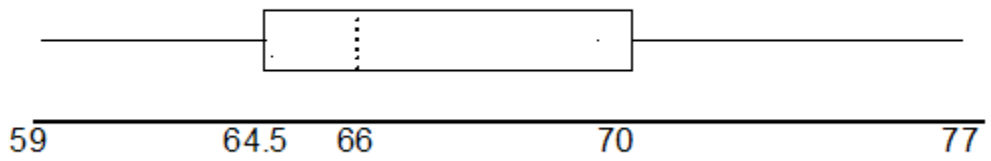
- Enter data into the list editor (Press STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, arrow down.
- Put the data values in list L1.
- Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.
- Press ENTER
- Use the down and up arrow keys to scroll.

- Smallest value = 59
- Largest value = 77
- Q1: First quartile = 64.5

- Q2: Second quartile or median= 66
- Q3: Third quartile = 70

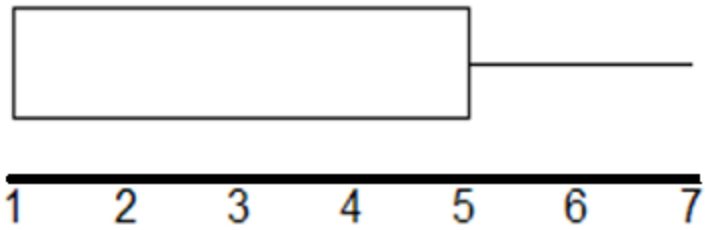
Using the TI-83, 83+, 84, 84+ to Construct the Box Plot
Go to 14:Appendix for Notes for the TI-83, 83+, 84, 84+ Calculator. To create the box plot:

- Press Y=. If there are any equations, press CLEAR to clear them.
- Press 2nd Y=.
- Press 4:Plotsoff. Press ENTER
- Press 2nd Y=
- Press 1:Plot1. Press ENTER.
- Arrow down and then use the right arrow key to go to the 5th picture which is the box plot. Press ENTER.
- Arrow down to Xlist: Press 2nd 1 for L1
- Arrow down to Freq: Press ALPHA. Press 1.
- Press ZOOM. Press 9:ZoomStat.
- Press TRACE and use the arrow keys to examine the box plot.



- **a**Each quarter has 25% of the data.
- **b**The spreads of the four quarters are 64.5 - 59 = 5.5 (first quarter), 66 - 64.5 = 1.5 (second quarter), 70 - 66 = 4 (3rd quarter), and 77 - 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
- **c**Interquartile Range: $\text{IQR} = \text{Q3} - \text{Q1} = 70 - 64.5 = 5.5$.
- **d**The interval 59 through 65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
- **e**The middle 50% (middle half) of the data has a range of 5.5 inches.

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both 1, the median and the third quartile were both 5, and the largest value was 7, the box plot would look as follows:



Example:
Test scores for a college statistics class held during the day are:
99 56 78 55.5 32 90 80 81 56 59 45 77 84.5 84 70 72 68 32 79 90
Test scores for a college statistics class held during the evening are:
98 78 68 83 81 89 88 76 65 45 98 90 80 84.5 85 79 78 98 90 79 81 25.5

Exercise:

Problem:

- What are the smallest and largest data values for each data set?
- What is the median, the first quartile, and the third quartile for each data set?
- Create a boxplot for each set of data.
- Which boxplot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

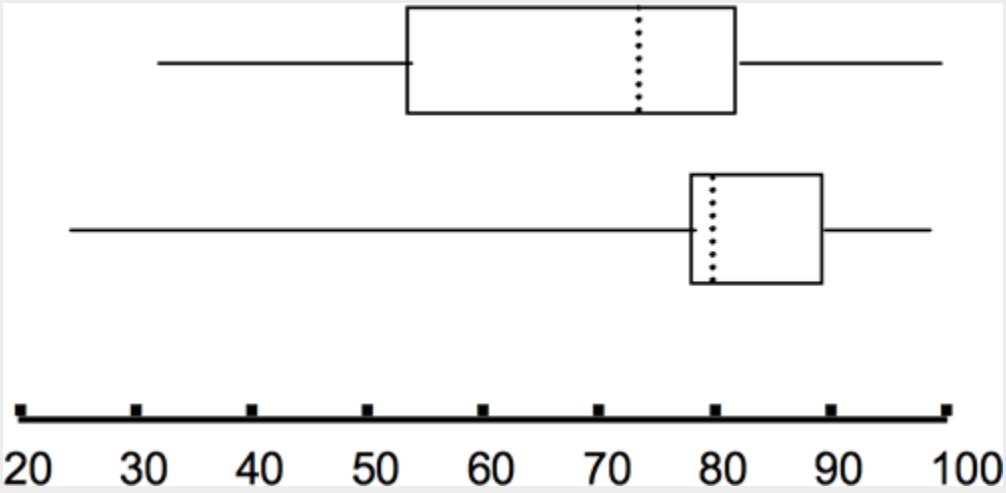
- For each data set, what percent of the data is between the smallest value and the first quartile? (Answer: 25%) the first quartile and the median? (Answer: 25%) the median and the third quartile? the third quartile and the largest value? What percent of the data is between the first quartile and the largest value? (Answer: 75%)

Solution:
First Data Set

- $\text{Xmin} = 32$
- $\text{Q1} = 56$
- $M = 74.5$
- $\text{Q3} = 82.5$
- $\text{Xmax} = 99$

Second Data Set

- $\text{Xmin} = 25.5$
- $\text{Q1} = 78$
- $M = 81$
- $\text{Q3} = 89$
- $\text{Xmax} = 98$



The first data set (the top box plot) has the widest spread for the middle 50% of the data. $\text{IQR} = \text{Q3} - \text{Q1}$ is $82.5 - 56 = 26.5$ for the first data set and $89 - 78 = 11$ for the second data set. So, the first set of data has its middle 50% of scores more spread out.
25% of the data is between M and Q3 and 25% is between Q3 and Xmax.

Glossary

Median
: A number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Quartiles
: The numbers that separate the data into quarters. Quartiles may or may not be part of the data. The second quartile is the median of the data.

Measures of the Location of the Data
Descriptive Statistics: Measuring the Location of Data explains percentiles and quartiles and is part of the collection col10555 written by Barbara Illowsky and Susan Dean. Roberta Bloom contributed the section "Interpreting Percentiles, Quartile and the Median."

The common measures of location are **quartiles** and **percentiles** (%iles). Quartiles are special percentiles. The first quartile, Q_1 is the same as the 25th percentile (25th %ile) and the third quartile, Q_3, is the same as the 75th percentile (75th %ile). The median, M, is called both the second quartile and the 50th percentile (50th %ile).

Note: Quartiles are given special attention in the Box Plots module in this chapter.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Recall that quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).
Equation:

$$\text{IQR} = Q_3 - Q_1$$

The IQR can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than $(1.5)(\text{IQR})$ below the first quartile or more than $(1.5)(\text{IQR})$ above the third quartile**. Potential outliers always need further investigation.

Example:
Exercise:

Problem:

For the following 13 real estate prices, calculate the IQR and determine if any prices are outliers. Prices are in dollars. (*Source: San Jose Mercury News*)

389,950 230,500 158,000 479,000 639,000 114,950 5,500,000 387,000 659,000 529,000 575,000 488,800 1,095,000

Solution:

Order the data from smallest to largest.

114,950 158,000 230,500 387,000 389,950 479,000 488,800 529,000 575,000 639,000 659,000 1,095,000 5,500,000

$M = 488{,}800$

$Q_1 = \frac{230500+387000}{2} = 308750$

$Q_3 = \frac{639000+659000}{2} = 649000$

$\text{IQR} = 649000 - 308750 = 340250$

$(1.5)(\text{IQR}) = (1.5)(340250) = 510375$

$Q_1 - (1.5)(\text{IQR}) = 308750 - 510375 = -201625$

$Q_3 + (1.5)(\text{IQR}) = 649000 + 510375 = 1159375$

No house price is less than -201625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Example:
Exercise:

Problem:

For the two data sets in the test scores example, find the following:

- **a**The interquartile range. Compare the two interquartile ranges.
- **b**Any outliers in either set.
- **c**The 30th percentile and the 80th percentile for each set. How much data falls below the 30th percentile? Above the 80th percentile?

Solution:

For the IQRs, see the answer to the test scores example. The first data set has the larger IQR, so the scores between $Q3$ and $Q1$ (middle 50%) for the first data set are more spread out and not clustered about the median.

First Data Set

- $\left(\frac{3}{2}\right) \cdot (\text{IQR}) = \left(\frac{3}{2}\right) \cdot (26.5) = 39.75$
- $\text{Xmax - Q3} = 99 - 82.5 = 16.5$
- $\text{Q1 - Xmin} = 56 - 32 = 24$

$\left(\frac{3}{2}\right) \cdot (\text{IQR}) = 39.75$ is larger than 16.5 and larger than 24, so the first set has no outliers.

Second Data Set

- $\left(\frac{3}{2}\right) \cdot (\text{IQR}) = \left(\frac{3}{2}\right) \cdot (11) = 16.5$
- $\text{Xmax} - \text{Q3} = 98 - 89 = 9$
- $\text{Q1} - \text{Xmin} = 78 - 25.5 = 52.5$

$\left(\frac{3}{2}\right) \cdot (\text{IQR}) = 16.5$ is larger than 9 but smaller than 52.5, so for the second set 45 and 25.5 are outliers.

To find the percentiles, create a frequency, relative frequency, and cumulative relative frequency chart (see "Frequency" from the Sampling and

Data Chapter). Get the percentiles from that chart.

First Data Set

- 30th %ile (between the 6th and 7th values) $= \frac{(56 + 59)}{2} = 57.5$
- 80th %ile (between the 16th and 17th values) $= \frac{(84 + 84.5)}{2} = 84.25$

Second Data Set

- 30th %ile (7th value) $= 78$
- 80th %ile (18th value) $= 90$

30% of the data falls below the 30th %ile, and 20% falls above the 80th %ile.

Example:
Finding Quartiles and Percentiles Using a Table
Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were (student data):

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Find the 28th percentile: Notice the 0.28 in the "cumulative relative frequency" column. 28% of 50 data values = 14. There are 14 values less than the 28th %ile. They include the two 4s, the five 5s, and the seven 6s. The 28th %ile is between the last 6 and the first 7. **The 28th %ile is 6.5.**
Find the median: Look again at the "cumulative relative frequency " column and find 0.52. The median is the 50th %ile or the second quartile. 50% of 50 = 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th %ile is between the 25th (7) and 26th (7) values. **The median is 7.**
Find the third quartile: The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the 4s, 5s, 6s and 7s, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th %ile, then, must be an 8** . Another way to look at the problem is to find 75% of 50 (= 37.5) and round up to 38. The third quartile, Q_3, is the 38th value which is an 8. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Example:

Exercise:

Problem: Using the table:

1. Find the 80th percentile.
2. Find the 90th percentile.
3. Find the first quartile.
4. What is another name for the first quartile?

Solution:

1. $\frac{(8+9)}{2} = 8.5$
   Look where cum. rel. freq. = 0.80. 80% of the data is 8 or less. 80th %ile is between the last 8 and first 9.
2. 9
3. 6
4. First Quartile = 25th %ile

Collaborative Classroom Exercise: Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions.

1. How many students were surveyed?
2. What kind of sampling did you do?
3. Construct a table of the data.
4. Construct 2 different histograms. For each, starting value = _____ ending value = ____.
5. Use the table to find the median, first quartile, and third quartile.
6. Construct a box plot.
7. Use the table to find the following:
   - The 10th percentile
   - The 70th percentile
   - The percent of students who own less than 4 sweaters

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order, from smallest to largest. p% of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad". The interpretation of whether a certain percentile is good or bad depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good'; in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to properly interpret percentiles is important not only when describing data, but is also important in later chapters of this textbook when calculating probabilities.

Guideline:

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information:

- information about the context of the situation being considered,
- the data value (value of the variable) that represents the percentile,
- the percent of individuals or items with data values below the percentile.
- Additionally, you may also choose to state the percent of individuals or items with data values above the percentile.

Example:
On a timed math test, the first quartile for times for finishing the exam was 35 minutes. Interpret the first quartile in the context of this situation.

- 25% of students finished the exam in 35 minutes or less.
- 75% of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Example:
On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

- 70% of students answered 16 or fewer questions correctly.
- 30% of students answered 16 or more questions correctly.
- Note: A high percentile could be considered good, as answering more questions correctly is desirable.

Example:
At a certain community college, it was found that the 30th percentile of credit units that students are enrolled for is 7 units. Interpret the 30th percentile in the context of this situation.

- 30% of students are enrolled in 7 or fewer credit units
- 70% of students are enrolled in 7 or more credit units
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Do the following Practice Problems for Interpreting Percentiles
Exercise:

Problem:

- **a** For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
- **b** The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.

- **c** A bicyclist in the 90th percentile of a bicycle race between two towns completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Solution:

- **a** For runners in a race it is more desirable to have a low percentile for finish time. A low percentile means a short time, which is faster.
- **b**INTERPRETATION: 20% of runners finished the race in 5.2 minutes or less. 80% of runners finished the race in 5.2 minutes or longer.
- **c**He is among the slowest cyclists (90% of cyclists were faster than him.) INTERPRETATION: 90% of cyclists had a finish time of 1 hour, 12 minutes or less.Only 10% of cyclists had a finish time of 1 hour, 12 minutes or longer

Exercise:

Problem:

- **a** For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
- **b**The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Solution:

- **a**For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed, which is faster.
- **b**INTERPRETATION: 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Exercise:

Problem:

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Solution:

On an exam you would prefer a high percentile; higher percentiles correspond to higher grades on the exam.

Exercise:

Problem:

Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Solution:

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than you did. In this context, you would prefer a wait time corresponding to a lower percentile. INTERPRETATION: 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

Exercise:

Problem:

In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

Solution:

Li should be pleased. Her salary is relatively high compared to other recent college grads. 78% of recent college graduates earn less than Li does. 22% of recent college graduates earn more than Li does.

Exercise:

Problem:

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1700 in damage and was in the 90th percentile. Should the manufacturer and/or a consumer be pleased or upset by this result? Explain. Write a sentence that interprets the 90th percentile in the context of this problem.

Solution:

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of $1700 or less; only 10% had damage repair costs of $1700 or more.

Exercise:

Problem:

- The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:
- a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
- b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percent of students from each high school are "eligible in the local context"?

Solution:

- **a**The top 12% of students are those who are at or above the **88th percentile** of admissions index scores.
- **b** The **top 4%** of students' GPAs are at or above the 96th percentile, making the top 4% of students "eligible in the local context".

Exercise:

Problem:

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Solution:

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost $240,000 or less. 66% of houses cost $240,000 or more.

**With contributions from Roberta Bloom

Glossary

Interquartile Range (IRQ)
: The distance between the third quartile (Q3) and the first quartile (Q1). IQR = Q3 - Q1.

Outlier
: An observation that does not fit the rest of the data.

Percentile
: A number that divides ordered data into hundredths.

Example:
Let a data set contain 200 ordered observations starting with $\{2.3, 2.7, 2.8, 2.9, 2.9, 3.0 \ldots\}$. Then the first percentile is $\frac{(2.7+2.8)}{2} = 2.75$, because 1% of the data is to the left of this point on the number line and 99% of the data is on its right. The second percentile is $\frac{(2.9+2.9)}{2} = 2.9$. Percentiles may or may not be part of the data. In this example, the first percentile is not in the data, but the second percentile is. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles

The numbers that separate the data into quarters. Quartiles may or may not be part of the data. The second quartile is the median of the data.

Measures of the Center of the Data
This chapter discusses measuring descriptive statistical information using the center of the data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts (previously discussed under box plots in this chapter). The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

Note: The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean" and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

The mean can also be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the sample mean is an x with a bar over it (pronounced "x bar"): $\overline{x}$.

The Greek letter μ (pronounced "mew") represents the population mean. One of the requirements for the sample mean to be a good estimate of the population mean is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:

11122344444
Equation:

$$\overline{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7$$

Equation:

$$\overline{x} = \frac{3 \times 1 + 2 \times 2 + 1 \times 3 + 5 \times 4}{11} = 2.7$$

In the second calculation for the sample mean, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by 2 after the data has been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2} = 49$. The median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$. The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example:
Exercise:

Problem:

AIDS data indicating the number of months an AIDS patient lives after taking a new antibody drug are as follows (smallest to largest):

3488101112131415151616171718212222242425262627272929313233334343534740444447

Calculate the mean and the median.

Solution:

The calculation for the mean is:

$$\overline{x} = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+...+35+37+40+(44)(2)+47]}{40} = 23.6$$

To find the median, **M**, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

34881011121314151516161717182122222424
25262627272929313233333434353740444447

$$M = \frac{24+24}{2} = 24$$

The median is 24.

Using the TI-83,83+,84, 84+ Calculators
Calculator Instructions are located in the menu item 14:Appendix (Notes for the TI-83, 83+, 84, 84+ Calculators).

- Enter data into the list editor. Press STAT 1:EDIT
- Put the data values in list L1.
- Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and ENTER.
- Press the down and up arrow keys to scroll.

$\overline{x} = 23.6$, $M = 24$

Example:
Exercise:

Problem:

Suppose that, in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center," the mean or the median?

Solution:

$\overline{x} = \frac{5000000+49\times30000}{50} = 129400$

$M = 30000$

(There are 49 people who earn $30,000 and one person who earns $5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The **mode** is the most frequent value. If a data set has two values that occur the same number of times, then the set is bimodal.

Example:
Statistics exam scores for 20 students are as follows
Statistics exam scores for 20 students are as follows:
50 53 59 59 63 63 72 72 72 72 72 76 78 81 83 84 84 84 90 93
Exercise:

Problem: Find the mode.

Solution:

The most frequent score is 72, which occurs five times. Mode = 72.

Example:
Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

Note: The mode can be calculated for qualitative data as well as for quantitative data.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean $\overline{x}$ of the sample is very likely to get closer and closer to μ. This is discussed in more detail in **The Central Limit Theorem**.

Note: The formula for the mean is located in the Summary of Formulas section course.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a **sampling distribution** as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	5/30
1	15/30
2	6/30
3	4/30
4	1/30

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A **statistic** is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean $\overline{x}$ is an example of a statistic which estimates the population mean μ.

Glossary

Mean
: A number that measures the central tendency. A common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by $\overline{x}$) is $\overline{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Median
: A number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

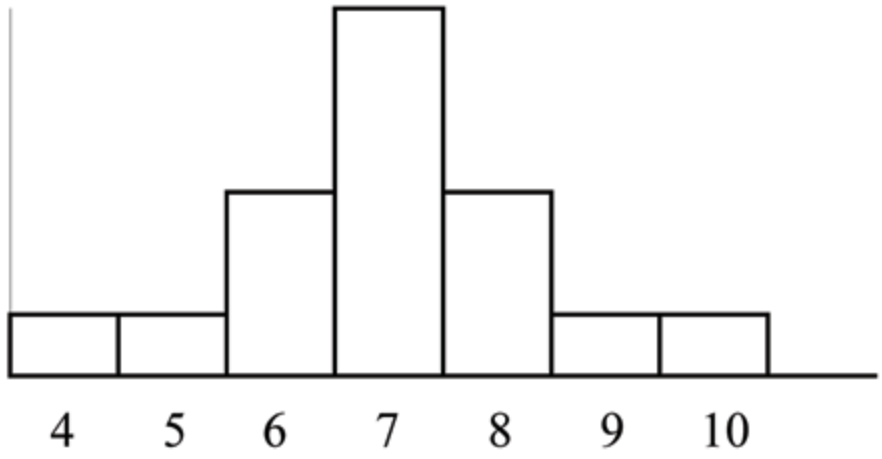
Mode
: The value that appears most frequently in a set of data.

Skewness and the Mean, Median, and Mode

Consider the following data set:

4 5 6 6 6 7 7 7 7 7 7 8 8 8 9 10

This data set produces the histogram shown below. Each interval has width one and each value is located in the middle of an interval.

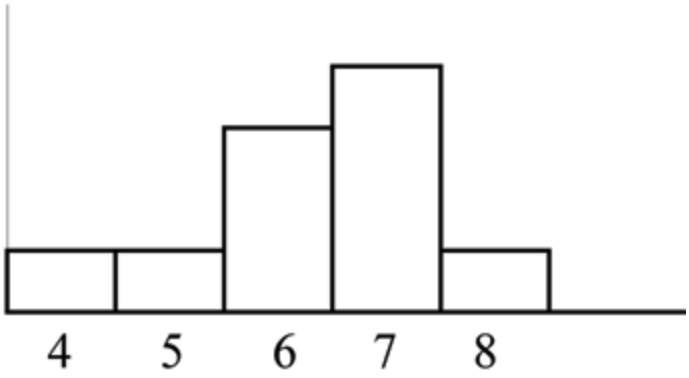


The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each 7 for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal) and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data:

4 5 6 6 6 7 7 7 7 8

is not symmetrical. The right-hand side seems "chopped off" compared to the left side. The shape distribution is called **skewed to the left** because it is pulled out to the left.

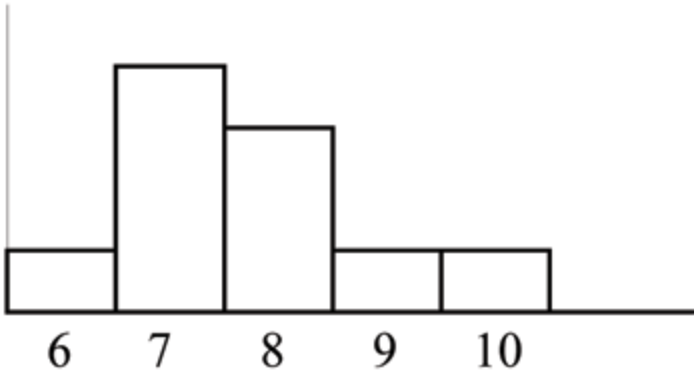


The mean is 6.3, the median is 6.5, and the mode is 7. **Notice that the mean is less than the median and they are both less than the mode.** The mean and the median both reflect the skewing but the mean more so.

The histogram for the data:

6 7 7 7 7 8 8 8 9 10

is also not symmetrical. It is **skewed to the right**.



The mean is 7.7, the median is 7.5, and the mode is 7. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Measures of the Spread of the Data

Descriptive Statistics: Measuring the Spread of Data explains standard deviation as a measure of variation in data and is part of the collection col10555 written by Barbara Illowsky and Susan Dean. Roberta Bloom made contributions that helped to clarify the standard deviation and the variance.

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation.

The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set
- can be used to determine whether a particular data value is close to or far from the mean

The standard deviation provides a measure of the overall variation in a data set
The standard deviation is always positive or 0. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying waiting times at the checkout line for customers at supermarket A and supermarket B; the average wait time at both markets is 5 minutes. At market A, the standard deviation for the waiting time is 2 minutes; at market B the standard deviation for the waiting time is 4 minutes.

Because market B has a higher standard deviation, we know that there is more variation in the waiting times at market B. Overall, wait times at market B are more spread out from the average; wait times at market A are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.
Suppose that Rosa and Binh both shop at Market A. Rosa waits for 7 minutes and Binh waits for 1 minute at the checkout counter. At market A, the mean wait time is 5 minutes and the standard deviation is 2 minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for 7 minutes:

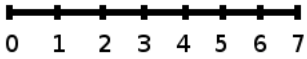
- 7 is 2 minutes longer than the average of 5; 2 minutes is equal to one standard deviation.
- Rosa's wait time of 7 minutes is **2 minutes longer than the average** of 5 minutes.
- Rosa's wait time of 7 minutes is **one standard deviation above the average** of 5 minutes.

Binh waits for 1 minute.

- 1 is 4 minutes less than the average of 5; 4 minutes is equal to two standard deviations.
- Binh's wait time of 1 minute is **4 minutes less than the average** of 5 minutes.
- Binh's wait time of 1 minute is **two standard deviations below the average** of 5 minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than 2 standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than 2 standard deviations. (We will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put 5 and 7 on a number line, 7 is to the right of 5. We say, then, that 7 is **one** standard deviation to the **right** of 5 because
$5 + (1)(2) = 7$.

If 1 were also part of the data set, then 1 is **two** standard deviations to the **left** of 5 because
$5 + (-2)(2) = 1$.



- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- 7 is **one standard deviation more than the mean** of 5 because: 7=5+**(1)**(2)
- 1 is **two standard deviations less than the mean** of 5 because: 1=5+**(−2)**(2)

The equation **value = mean + (#ofSTDEVs)(standard deviation)** can be expressed for a sample and for a population:

- **sample:** $x = x + (\#\text{ofSTDEV})(\text{s})$
- **Population:** $x = \mu + (\#\text{ofSTDEV})(\sigma)$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol x is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation
If x is a number, then the difference "x - mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - x$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The **variance** is an **average of the squares of the deviations** (the $x - x$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by **N**, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by **n-1**, one less than the number of items in the sample. You can see that in the formulas below.

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x-x)^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f \cdot (x-x)^2}{n-1}}$
- For the sample standard deviation, the denominator is **n-1**, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

- $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f \cdot (x-\mu)^2}{N}}$
- For the population standard deviation, the denominator is **N**, the number of items in the population.

In these formulas, f represents the frequency with which a value appears. For example, if a value appears once, f is 1. If a value appears three times in the data set or population, f is 3.

Sampling Variability of a Statistic
The statistic of a sampling distribution was discussed in **Descriptive Statistics: Measuring the Center of the Data**. How much the statistic varies from one sample to another is known as the **sampling variability of a statistic**. You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in **The Central Limit Theorem** (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

Note: In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83,83+,84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean.

Example:
In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:
9 9.5 9.5 10 10 10 10 10.5 10.5 10.5 10.5 11 11 11 11 11 11 11.5 11.5 11.5
Equation:

$$x = \frac{9 + 9.5 \times 2 + 10 \times 4 + 10.5 \times 4 + 11 \times 6 + 11.5 \times 3}{20} = 10.525$$

The average age is 10.53 years, rounded to 2 places.
The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s.

Data	**Freq.**	**Deviations**	Deviations^2	**(Freq.)(**Deviations^2**)**
x	f	$(x - x)$	$(x - x)^2$	$(f)(x - x)^2$
9	1	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$	$4 \times .275625 = 1.1025$
10.5	4	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$	$4 \times .000625 = .0025$
11	6	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$	$6 \times .225625 = 1.35375$

Data	Freq.	Deviations	Deviations^2	(Freq.)(Deviations^2)
11.5	3	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$	$3 \times .950625 = 2.851875$

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 - 1):
$s^2 = \frac{9.7375}{20-1} = 0.5125$
The **sample standard deviation** s is equal to the square root of the sample variance:
$s = \sqrt{0.5125} = .0715891$ Rounded to two decimal places, $s = 0.72$
Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.
Exercise:

Problem: Verify the mean and standard deviation calculated above on your calculator or computer.

Solution:
Using the TI-83,83+,84+ Calculators

- Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.
- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
- x=10.525
- Use Sx because this is sample data (not a population): Sx=0.715891

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**
- For a sample: x = x + (#ofSTDEVs)(s)
- For a population: x = μ + (#ofSTDEVs)(σ)
- For this example, use x = x + (#ofSTDEVs)(s) because the data is from a sample

Exercise:

Problem: Find the value that is 1 standard deviation above the mean. Find $(x + 1s)$.

Solution:

$(x + 1s) = 10.53 + (1)(0.72) = 11.25$

Exercise:

Problem: Find the value that is two standard deviations below the mean. Find $(x - 2s)$.

Solution:

$(x - 2s) = 10.53 - (2)(0.72) = 9.09$

Exercise:

Problem: Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution:

- $(x - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$
- $(x + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

Explanation of the standard deviation calculation shown in the table
The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11. The deviations 0.97 and 0.47 indicate that. A positive deviation occurs when the data value is greater than the mean. A negative deviation occurs when the data value is less than the mean; the deviation is -1.525 for the data value 9. **If you add the deviations, the sum is always zero**. (For this example, there are n=20 deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by n=20, the calculation divided by n-1=20-1=19 because the data is a sample. For the **sample** variance, we divide by the sample size minus one (n-1). Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by (n-1) gives a better estimate of the population variance.

Note: Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is 0, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**.

Note: The formula for the standard deviation is at the end of the chapter.

Example:
Exercise:

Problem: Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

3342494953555561 6367686869697273 7478808388888890 929494949496100

- **a** Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
- **b** Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
    - **i** The sample mean
    - **ii** The sample standard deviation
    - **iii** The median
    - **iv** The first quartile
    - **v** The third quartile
    - **vi** IQR
- **c** Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

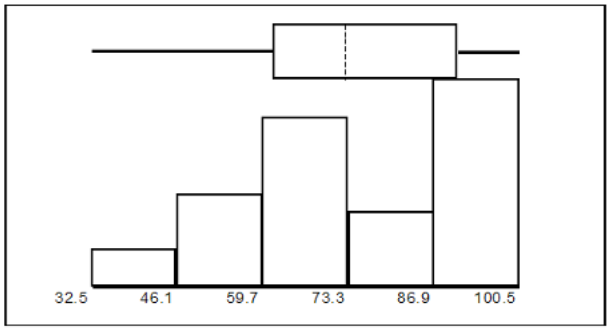
Solution:

- **a**

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	**0.998** (Why isn't this value 1?)

- **b**
  - **i**The sample mean = 73.5
  - **ii**The sample standard deviation = 17.9
  - **iii**The median = 73
  - **iv**The first quartile = 61
  - **v**The third quartile = 90
  - **vi**IQR = 90 - 61 = 29
- **c**The x-axis goes from 32.5 to 100.5; y-axis goes from -2.4 to 15 for the histogram; number of intervals is 5 for the histogram so the width of an interval is (100.5 - 32.5) divided by 5 which is equal to 13.6. Endpoints of the intervals: starting point is 32.5, 32.5+13.6 = 46.1, 46.1+13.6 = 59.7, 59.7+13.6 = 73.3, 73.3+13.6 = 86.9, 86.9+13.6 = 100.5 = the ending value; No data values fall on an interval boundary.



The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater (73 - 33 = 40) than the spread in the upper 50% (100 - 73 = 27). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (IQR = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, it can be misleading to compare the data values directly.

- For each data value, calculate how many standard deviations the value is away from its mean.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\#\text{ofSTDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z. In symbols, the formulas become:

Sample	$x = \bar{x} + z\, s$	$z = \frac{x - \bar{x}}{s}$
Population	$x = \mu + z\, \sigma$	$z = \frac{x - \mu}{\sigma}$

Example:
Exercise:

Problem:

Two students, John and Ali, from different high schools, wanted to find out who had the highest G.P.A. when compared to his school. Which student had the highest G.P.A. when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Solution:

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$\#\text{ofSTDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$; $z = \frac{x - \mu}{\sigma}$

For John, $z = \#\text{ofSTDEVs} = \frac{2.85 - 3.0}{0.7} = -0.21$

For Ali, $z = \#\text{ofSTDEVs} = \frac{77 - 80}{10} = -0.3$

John has the better G.P.A. when compared to his school because his G.P.A. is 0.21 standard deviations **below** his school's mean while Ali's G.P.A. is 0.3 standard deviations **below** his school's mean.

John's z-score of -0.21 is higher than Ali's z-score of -0.3 . For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within 2 standard deviations of the mean.
- At least 89% of the data is within 3 standard deviations of the mean.
- At least 95% of the data is within 4 1/2 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is MOUND-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within 1 standard deviation of the mean.
- Approximately 95% of the data is within 2 standard deviations of the mean.
- More than 99% of the data is within 3 standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is mound-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

**With contributions from Roberta Bloom

Glossary

Standard Deviation
: A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Variance
: Mean of the squared deviations from the mean. Square of the standard deviation. For a set of data, a deviation can be represented as $x - x$ where x is a value of the data and x is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and 1.

Summary of Formulas
A summary of useful formulas used in examining descriptive statistics

Commonly Used Symbols

- The symbol Σ means to add or to find the sum.
- n = the number of data values in a sample
- N = the number of people, things, etc. in the population
- x = the sample mean
- s = the sample standard deviation
- μ = the population mean
- σ = the population standard deviation
- f = frequency
- x = numerical value

Commonly Used Expressions

- x^*f = A value multiplied by its respective frequency
- $\sum x$ = The sum of the values
- $\sum x^*f$ = The sum of values multiplied by their respective frequencies
- $(x - x)$ or $(x - \mu)$ = Deviations from the mean (how far a value is from the mean)
- $(x - x)^2$ or $(x - \mu)^2$ = Deviations squared
- $f(x - x)^2$ or $f(x - \mu)^2$ = The deviations squared and multiplied by their frequencies

Mean Formulas:

- $x = \frac{\sum x}{n}$ or $x = \frac{\sum f \cdot x}{n}$
- $\mu = \frac{\sum x}{N}$ or $\mu = \frac{\sum f \cdot x}{N}$

Standard Deviation Formulas:

- $s = \sqrt{\frac{\Sigma(x-x)^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f \cdot (x-x)^2}{n-1}}$
- $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f \cdot (x-\mu)^2}{N}}$

Formulas Relating a Value, the Mean, and the Standard Deviation:

- value = mean + (#ofSTDEVs)(standard deviation), where #ofSTDEVs = the number of standard deviations
- $x = x+$ (#ofSTDEVs)(s)
- $x = \mu$ + (#ofSTDEVs)(σ)

Practice 1: Center of the Data
This module provides students with opportunities to apply concepts related to descriptive statistics. Students are asked to take a set of sample data and calculate a series of statistical values for that data.

Student Learning Outcomes

- The student will calculate and interpret the center, spread, and location of the data.
- The student will construct and interpret histograms an box plots.

Given

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

Complete the Table

Data Value (# cars)	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**

Discussion Questions

Exercise:

Problem: What does the frequency column sum to? Why?

Solution:

65

Exercise:

Problem: What does the relative frequency column sum to? Why?

Solution:

1

Exercise:

Problem:

What is the difference between relative frequency and frequency for each data value?

Exercise:

Problem:

What is the difference between cumulative relative frequency and relative frequency for each data value?

Enter the Data

Enter your data into your calculator or computer.

Construct a Histogram

Determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram below. Label the horizontal and vertical axes with words. Include numerical scaling.

Data Statistics

Calculate the following values:

Exercise:

Problem: Sample mean = x =

Solution:

4.75

Exercise:

Problem: Sample standard deviation = s_x =

Solution:

1.39

Exercise:

Problem: Sample size = n =

Solution:

65

Calculations

Use the table in section 2.11.3 to calculate the following values:

Exercise:

Problem: Median =

Solution:

4

Exercise:

Problem: Mode =

Solution:

4

Exercise:

Problem: First quartile =

Solution:

4

Exercise:

Problem: Second quartile = median = 50th percentile =

Solution:

4

Exercise:

Problem: Third quartile =

Solution:

6

Exercise:

Problem: Interquartile range ( ) = _____ - _____ = _____

Solution:

Exercise:

Problem: 10th percentile =

Solution:

3

Exercise:

Problem: 70th percentile =

Solution:

6

Exercise:

Problem: Find the value that is 3 standard deviations:

- **a**Above the mean
- **b**Below the mean

Solution:

- **a**8.93

- **b**0.58

Box Plot

Construct a box plot below. Use a ruler to measure and scale accurately.

Interpretation

Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

Practice 2: Spread of the Data
Practice exercise for Descriptive Statistics

Student Learning Outcomes

- The student will calculate measures of the center of the data.
- The student will calculate the spread of the data.

Given

The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976-77 through 2004-2005. (*Source: Graphically Speaking by Bill King, LTCC Institutional Research, December 2005*).

Use these values to answer the following questions:

- μ = 1000 FTES
- Median = 1014 FTES
- σ = 474 FTES
- First quartile = 528.5 FTES
- Third quartile = 1447.5 FTES
- n = 29 years

Calculate the Values

Exercise:

Problem:

A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

Solution:

6

Exercise:

Problem: 75% of all years have a FTES:

- **a**At or below:
- **b**At or above:

Solution:

- **a**1447.5
- **b**528.5

Exercise:

Problem: The population standard deviation =

Solution:

474 FTES

Exercise:

Problem:

What percent of the FTES were from 528.5 to 1447.5? How do you know?

Solution:

50%

Exercise:

Problem: What is the IQR? What does the IQR represent?

Solution:

919

Exercise:

Problem:

How many standard deviations away from the mean is the median?

Solution:

0.03

Additional Information: The population FTES for 2005-2006 through 2010-2011 was given in an updated report. (Source: http://www.ltcc.edu/data/ResourcePDF/LTCC_FactBook_2010-11.pdf). The data are reported here.

Year	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11
Total FTES	1585	1690	1735	1935	2021	1890

Exercise:

Problem:

Calculate the mean, median, standard deviation, first quartile, the third quartile and the IQR. Round to one decimal place.

Solution:

mean = 1809.3
median = 1812.5
standard deviation = 151.2
First quartile = 1690

Third quartile = 1935
IQR = 245

Exercise:

Problem:

Construct a boxplot for the FTES for 2005-2006 through 2010-2011 and a boxplot for the FTES for 1976-1977 through 2004-2005.

Exercise:

Problem:

Compare the IQR for the FTES for 1976-77 through 2004-2005 with the IQR for the FTES for 2005-2006 through 2010-2011. Why do you suppose the IQRs are so different?

Solution:

Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

Homework

Descriptive Statistics: Homework is part of the collection col10555 written by Barbara Illowsky and Susan Dean and provides homework questions related to lessons about descriptive statistics.

Exercise:

Problem:

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

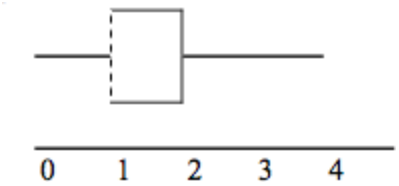
# of movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

- **a**Find the sample mean x
- **b**Find the sample standard deviation, s
- **c**Construct a histogram of the data.
- **d**Complete the columns of the chart.
- **e**Find the first quartile.
- **f**Find the median.
- **g**Find the third quartile.
- **h**Construct a box plot of the data.

- **i**What percent of the students saw fewer than three movies?
- **j**Find the 40th percentile.
- **k**Find the 90th percentile.
- **l**Construct a line graph of the data.
- **m**Construct a stem plot of the data.

Solution:

- **a**1.48
- **b**1.12
- **e**1
- **f**1
- **g**2
- **h**



- **i**80%
- **j**1
- **k**3

Exercise:

Problem:

The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years. ((*Source: http://www.usatoday.com/news/nation/story/2012-05-17/minority-births-census/55029100/1)*)

- **a**Based upon this information, give two reasons why the black median age could be lower than the white median age.
- **b**Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?

- **c**How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

Exercise:

Problem:

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

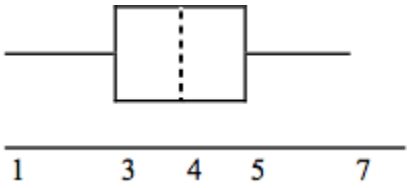
X	Frequency	Relative Frequency	Cumulative Relative Frequency
1	2		
2	5		
3	8		
4	12		
5	12		
7	1		

- **a**Find the sample mean x
- **b**Find the sample standard deviation, s
- **c**Construct a histogram of the data.
- **d**Complete the columns of the chart.
- **e**Find the first quartile.

- **f**Find the median.
- **g**Find the third quartile.
- **h**Construct a box plot of the data.
- **i**What percent of the students owned at least five pairs?
- **j**Find the 40th percentile.
- **k**Find the 90th percentile.
- **l**Construct a line graph of the data
- **m**Construct a stem plot of the data

Solution:

- **a**3.78
- **b**1.29
- **e**3
- **f**4
- **g**5
- **h**



- **i**32.5%
- **j**4
- **k**5

Exercise:

Problem:

600 adult Americans were asked by telephone poll, What do you think constitutes a middle-class income? The results are below. Also, include left endpoint, but not the right endpoint. (*Source: Time magazine; survey by Yankelovich Partners, Inc.*)

Note:"Not sure" answers were omitted from the results.

Salary ($)	Relative Frequency
< 20,000	0.02
20,000 - 25,000	0.09
25,000 - 30,000	0.19
30,000 - 40,000	0.26
40,000 - 50,000	0.18
50,000 - 75,000	0.17
75,000 - 99,999	0.02
100,000+	0.01

- **a**What percent of the survey answered "not sure" ?
- **b**What percent think that middle-class is from $25,000 - $50,000 ?
- **c**Construct a histogram of the data

    1. **i**Should all bars have the same width, based on the data? Why or why not?
    2. **ii**How should the <20,000 and the 100,000+ intervals be handled? Why?

- **d**Find the 40th and 80th percentiles
- **e**Construct a bar graph of the data

Exercise:

Problem:

Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year (*Source: San Jose Mercury News*)

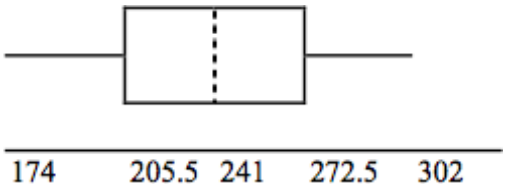
177 205 210 210 232 205 185 185 178 210 206 212 184 174 185 242 188 212 215 247 241 223 220 260 245 259 278 270 280 295 275 285 290 272 273 280 285 286 200 215 185 230 250 241 190 260 250 302 265 290 276 228 265

- **a** Organize the data from smallest to largest value.
- **b** Find the median.
- **c** Find the first quartile.
- **d** Find the third quartile.
- **e** Construct a box plot of the data.
- **f** The middle 50% of the weights are from _____ to _____.
- **g** If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
- **h** If our population were the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
- **i** Assume the population was the San Francisco 49ers. Find:
    - **i** the population mean, μ.
    - **ii** the population standard deviation, σ.
    - **iii** the weight that is 2 standard deviations below the mean.
    - **iv** When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?
- **j** That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who

was lighter, Smith or Young? How did you determine your answer?

Solution:

- **b** 241
- **c** 205.5
- **d** 272.5
- **e**



- **f** 205.5, 272.5
- **g** sample
- **h** population
- **i**
  - **i** 236.34
  - **ii** 37.50
  - **iii** 161.34
  - **iv** 0.84 std. dev. below the mean
- **j** Young

Exercise:

Problem:

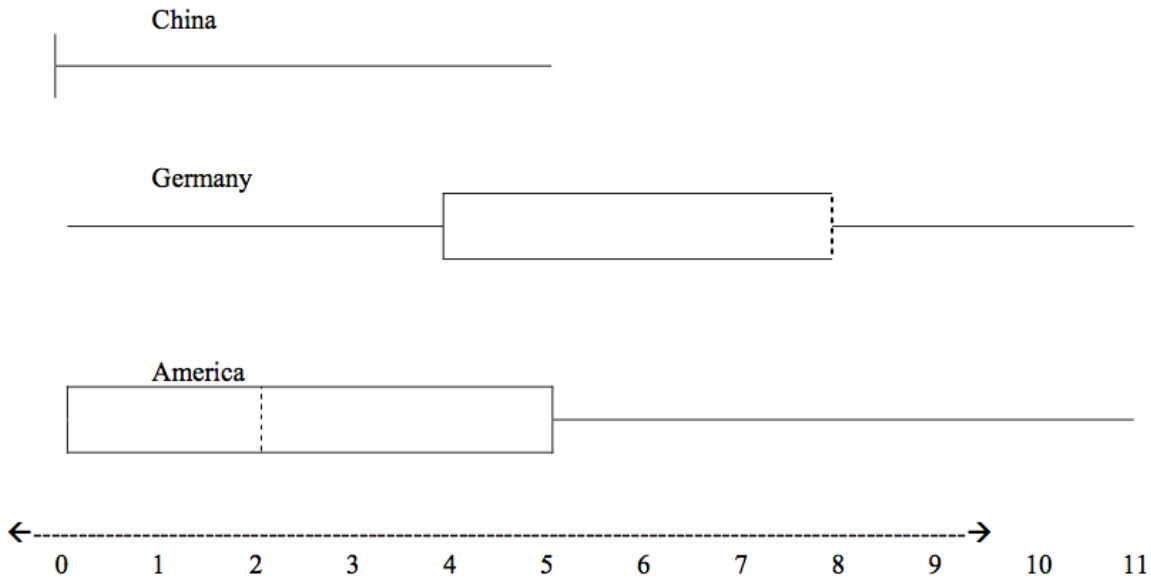
An elementary school class ran 1 mile with a mean of 11 minutes and a standard deviation of 3 minutes. Rachel, a student in the class, ran 1 mile in 8 minutes. A junior high school class ran 1 mile with a mean of 9 minutes and a standard deviation of 2 minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran 1 mile with a mean of 7 minutes and a standard deviation of 4 minutes. Nedda, a student in the class, ran 1 mile in 8 minutes.

- **a**Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
- **b**Who is the fastest runner with respect to his or her class? Explain why.

Exercise:

Problem:

In a survey of 20 year olds in China, Germany and America, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.



- **a**In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
- **b**Explain how it is possible that more Americans than Germans surveyed have been to over eight foreign countries.
- **c**Compare the three box plots. What do they imply about the foreign travel of twenty year old residents of the three countries when compared to each other?

Exercise:

Problem:

One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The twelve change scores are as follows:

3 8 -1 2 0 5 -3 1 -1 6 5 -2

- **a**What is the mean change score?
- **b**What is the standard deviation for this population?
- **c**What is the median change score?
- **d**Find the change score that is 2.2 standard deviations below the mean.

Exercise:

Problem:

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best G.P.A. when compared to his school? Explain how you determined your answer.

Student	G.P.A.	School Ave. G.P.A.	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Solution:

Kamala

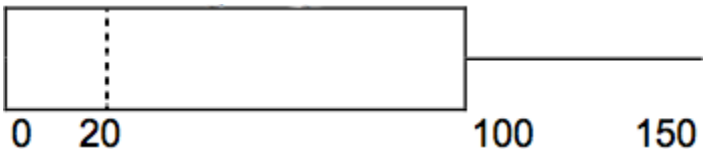
Exercise:

Problem: Given the following box plot:

- **a**Which quarter has the smallest spread of data? What is that spread?
- **b**Which quarter has the largest spread of data? What is that spread?
- **c**Find the Inter Quartile Range (IQR).
- **d**Are there more data in the interval 5 - 10 or in the interval 10 - 13? How do you know this?
- **e**Which interval has the fewest data in it? How do you know this?
    - **I** 0-2
    - **II**2-4
    - **III**10-12
    - **IV**12-13

Exercise:

Problem: Given the following box plot:



- **a**Think of an example (in words) where the data might fit into the above box plot. In 2-5 sentences, write down the example.
- **b**What does it mean to have the first and second quartiles so close together, while the second to fourth quartiles are far apart?

Exercise:

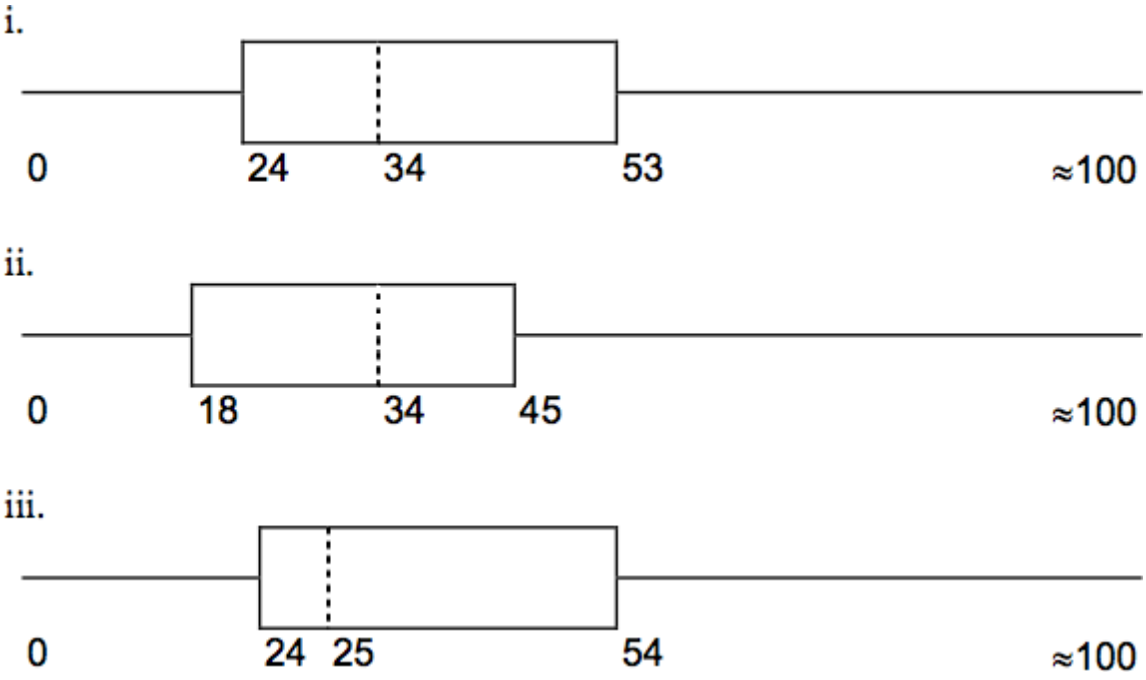
Problem:

Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows. (*Source: West magazine*)

Age Group	Percent of Community
0-17	18.9
18-24	8.0
25-34	22.8
35-44	15.0
45-54	13.1

Age Group	Percent of Community
55-64	11.9
65+	10.3

- **a**Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not?
- **b**What percent of the community is under age 35?
- **c**Which box plot most resembles the information above?



Exercise:

Problem:

Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, each asked adult consumers the number of fiction paperbacks they had purchased the previous month. The results are below.

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Publisher A

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	

# of books	Freq.	Rel. Freq.
5	10	
7	5	
9	1	

Publisher B

# of books	Freq.	Rel. Freq.
0-1	20	
2-3	35	
4-5	12	
6-7	2	
8-9	1	

Publisher C

- **a**Find the relative frequencies for each survey. Write them in the charts.
- **b**Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of 1. For Publisher C, make bar widths of 2.
- **c**In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.

- **d**Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
- **e**Make new histograms for Publisher A and Publisher B. This time, make bar widths of 2.
- **f**Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

Exercise:

Problem:

Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all on-board transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Below is a summary of the bills for each group.

Amount($)	Frequency	Rel. Frequency
51-100	5	
101-150	10	
151-200	15	
201-250	15	
251-300	10	
301-350	5	

Singles

Amount($)	Frequency	Rel. Frequency
100-150	5	
201-250	5	
251-300	5	
301-350	5	
351-400	10	
401-450	10	
451-500	10	
501-550	10	
551-600	5	
601-650	5	

Couples

- **a**Fill in the relative frequency for each group.
- **b**Construct a histogram for the Singles group. Scale the x-axis by $50. widths. Use relative frequency on the y-axis.
- **c**Construct a histogram for the Couples group. Scale the x-axis by $50. Use relative frequency on the y-axis.
- **d**Compare the two graphs:

    - **i**List two similarities between the graphs.
    - **ii**List two differences between the graphs.
    - **iii**Overall, are the graphs more similar or different?

- **e**Construct a new graph for the Couples by hand. Since each couple is paying for two individuals, instead of scaling the x-axis by $50, scale it by $100. Use relative frequency on the y-axis.
- **f**Compare the graph for the Singles with the new graph for the Couples:

    - **i**List two similarities between the graphs.
    - **ii**Overall, are the graphs more similar or different?

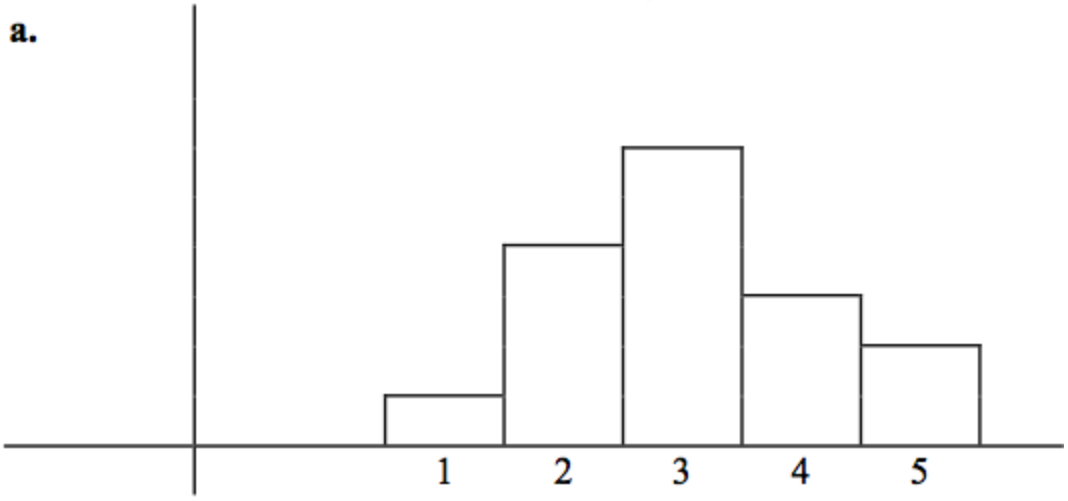
- **i**By scaling the Couples graph differently, how did it change the way you compared it to the Singles?
- **j**Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person in a couple? Explain why in one or two complete sentences.

Exercise:

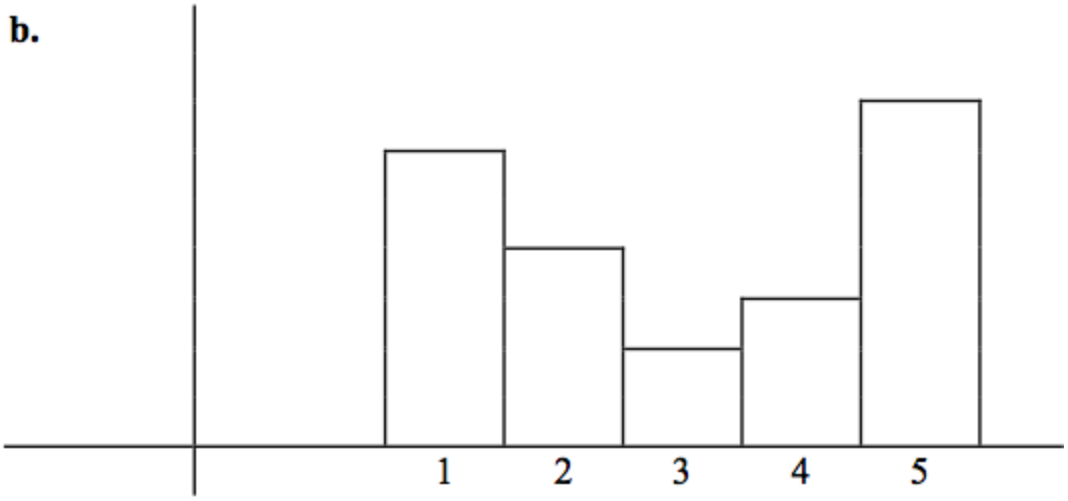
Problem:

Refer to the following histograms and box plot. Determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

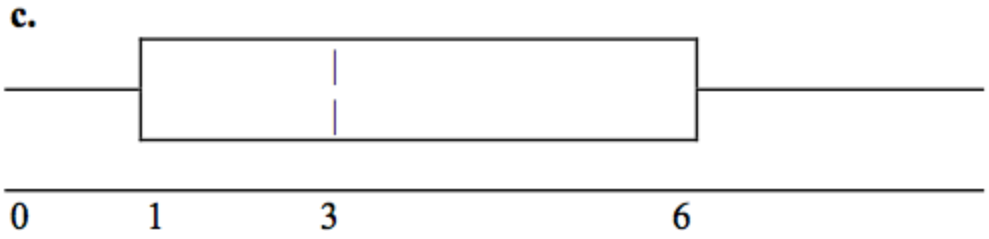
a.



b.



c.



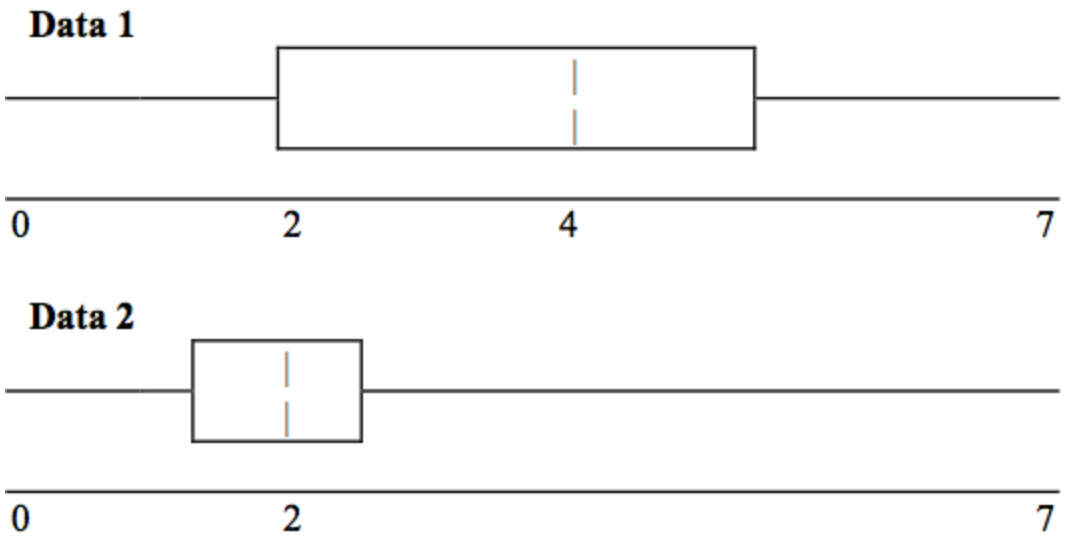
- **a**The medians for all three graphs are the same.
- **b**We cannot determine if any of the means for the three graphs is different.
- **c**The standard deviation for (b) is larger than the standard deviation for (a).
- **d**We cannot determine if any of the third quartiles for the three graphs is different.

Solution:

- **a**True
- **b**True
- **c**True
- **d**False

Exercise:

Problem: Refer to the following box plots.



- **a**In complete sentences, explain why each statement is false.
  - **iData 1** has more data values above 2 than **Data 2** has above 2.
  - **ii**The data sets cannot have the same mode.
  - **iii**For **Data 1**, there are more data values below 4 than there are above 4.
- **b**For which group, Data 1 or Data 2, is the value of “7” more likely to be an outlier? Explain why in complete sentences

Exercise:

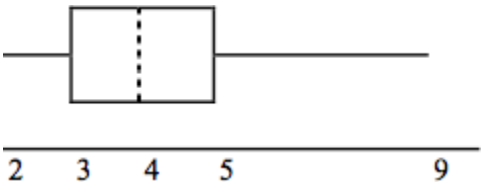
Problem:

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

- **a**Organize the data in a chart.
- **b**Find the median, the first quartile, and the third quartile.
- **c**Find the 65th percentile.
- **d**Find the 10th percentile.
- **e**Construct a box plot of the data.
- **f**The middle 50% of the conferences last from _____ days to _____ days.
- **g**Calculate the sample mean of days of engineering conferences.
- **h**Calculate the sample standard deviation of days of engineering conferences.
- **i**Find the mode.
- **j**If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
- **k**Give two reasons why you think that 3 - 5 days seem to be popular lengths of engineering conferences.

Solution:

- **b**4,3,5
- **c**4
- **d**3
- **e**



- **f**3,5
- **g**3.94
- **h**1.28
- **i**3
- **j**mode

Exercise:

Problem:

A survey of enrollment at 35 community colleges across the United States yielded the following figures (*source: Microsoft Bookshelf*):

6414 1550 2109 9350 21828 4300 5944 5722 2825 2044 5481 5200 5853 2750 10012 6357 27000 9414 7681 3200 17500 9200 7380 18314 6557 13713 17768 7493 2771 2861 1263 7285 28165 5080 11622

- **a**Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
- **b**Construct a histogram of the data.
- **c**If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
- **d**Calculate the sample mean.
- **e**Calculate the sample standard deviation.
- **f**A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Exercise:

Problem:

The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years. (*Source: Bureau of the Census*)

- **a**What does it mean for the median age to rise?
- **b**Give two reasons why the median age could rise.
- **c**For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

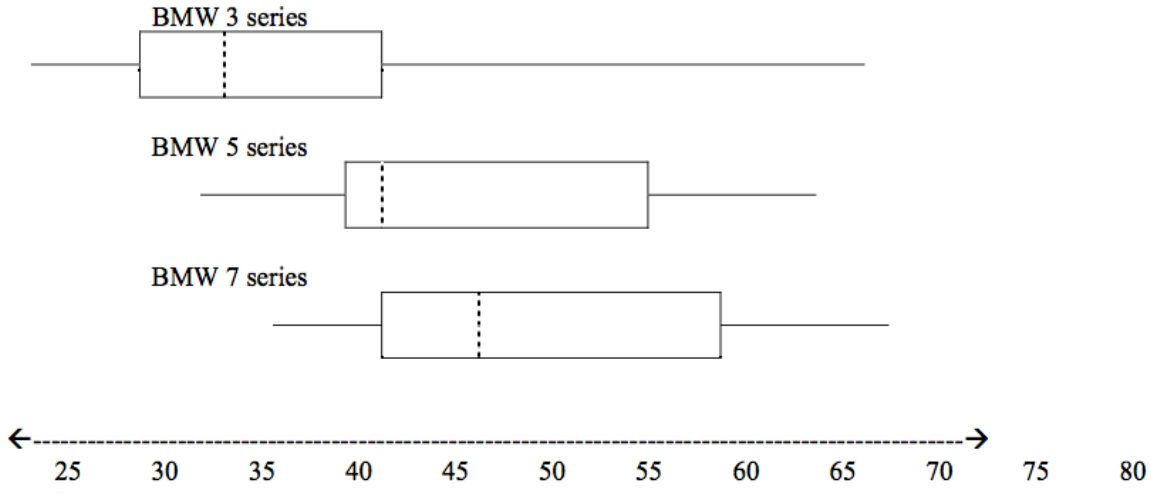
Solution:

- **c**Maybe

Exercise:

Problem:

A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.



- **a**In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car

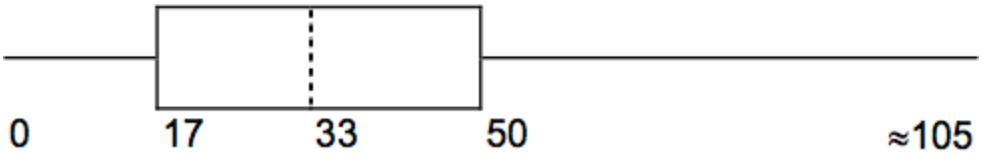
series.

- **b**Which group is most likely to have an outlier? Explain how you determined that.
- **c**Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?
- **d**Look at the BMW 5 series. Which quarter has the smallest spread of data? What is that spread?
- **e**Look at the BMW 5 series. Which quarter has the largest spread of data? What is that spread?
- **f**Look at the BMW 5 series. Estimate the Inter Quartile Range (IQR).
- **g**Look at the BMW 5 series. Are there more data in the interval 31-38 or in the interval 45-55? How do you know this?
- **h**Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?
    - **i**31-35
    - **ii**38-41
    - **iii**41-64

Exercise:

Problem:

The following box plot shows the U.S. population for 1990, the latest available year. (Source: Bureau of the Census, 1990 Census)



- **a**Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
- **b**12.6% are age 65 and over. Approximately what percent of the population are of working age adults (above age 17 to age 65)?

Solution:

- **a**more children
- **b**62.4%

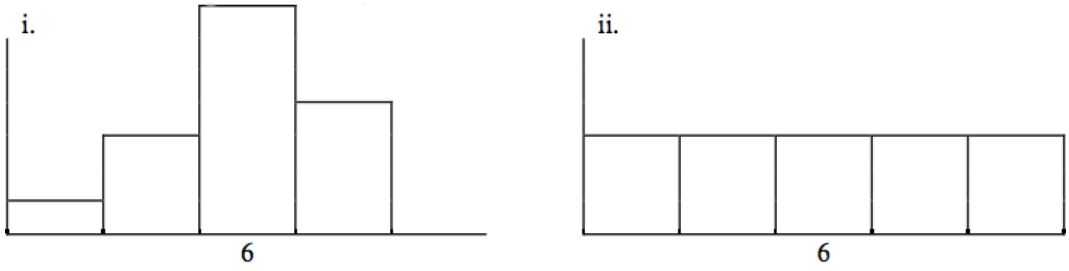
Exercise:

Problem:

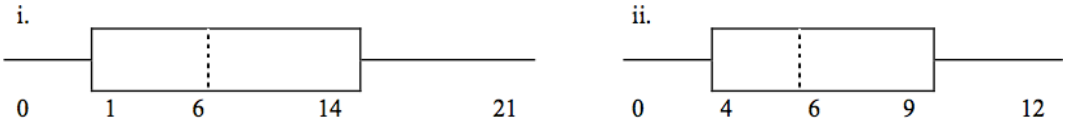
Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information:

	Javier	**Ercilla**
x	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

- **a**How can you determine which survey was correct ?
- **b**Explain what the difference in the results of the surveys implies about the data.
- **c**If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



- **d**If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia’s sample? How do you know?



Exercise:

Problem: Student grades on a chemistry exam were:

77, 78, 76, 81, 86, 51, 79, 82, 84, 99

- **a**Construct a stem-and-leaf plot of the data.
- **b**Are there any potential outliers? If so, which scores are they? Why do you consider them outliers?

Solution:

- **b**51,99

Try these multiple choice questions (Exercises 24 - 30).

The next three questions refer to the following information. We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency

Number of years	Frequency
7	1
14	3
15	1
18	1
19	4
20	3
22	1
23	1
26	1
40	2
42	2
	Total = 20

Exercise:

Problem: What is the IQR?

- **A**8
- **B**11
- **C**15
- **D**35

Solution:

A

Exercise:

Problem: What is the mode?

- **A**19
- **B**19.5
- **C**14 and 20
- **D**22.65

Solution:

A

Exercise:

Problem: Is this a sample or the entire population?

- **A**sample
- **B**entire population
- **C**neither

Solution:

B

The next two questions refer to the following table. X = the number of days per week that 100 clients use a particular exercise facility.

x	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

Exercise:

Problem: The 80th percentile is:

- **A**5
- **B**80
- **C**3
- **D**4

Solution:

D

Exercise:

Problem:

The number that is 1.5 standard deviations BELOW the mean is approximately:

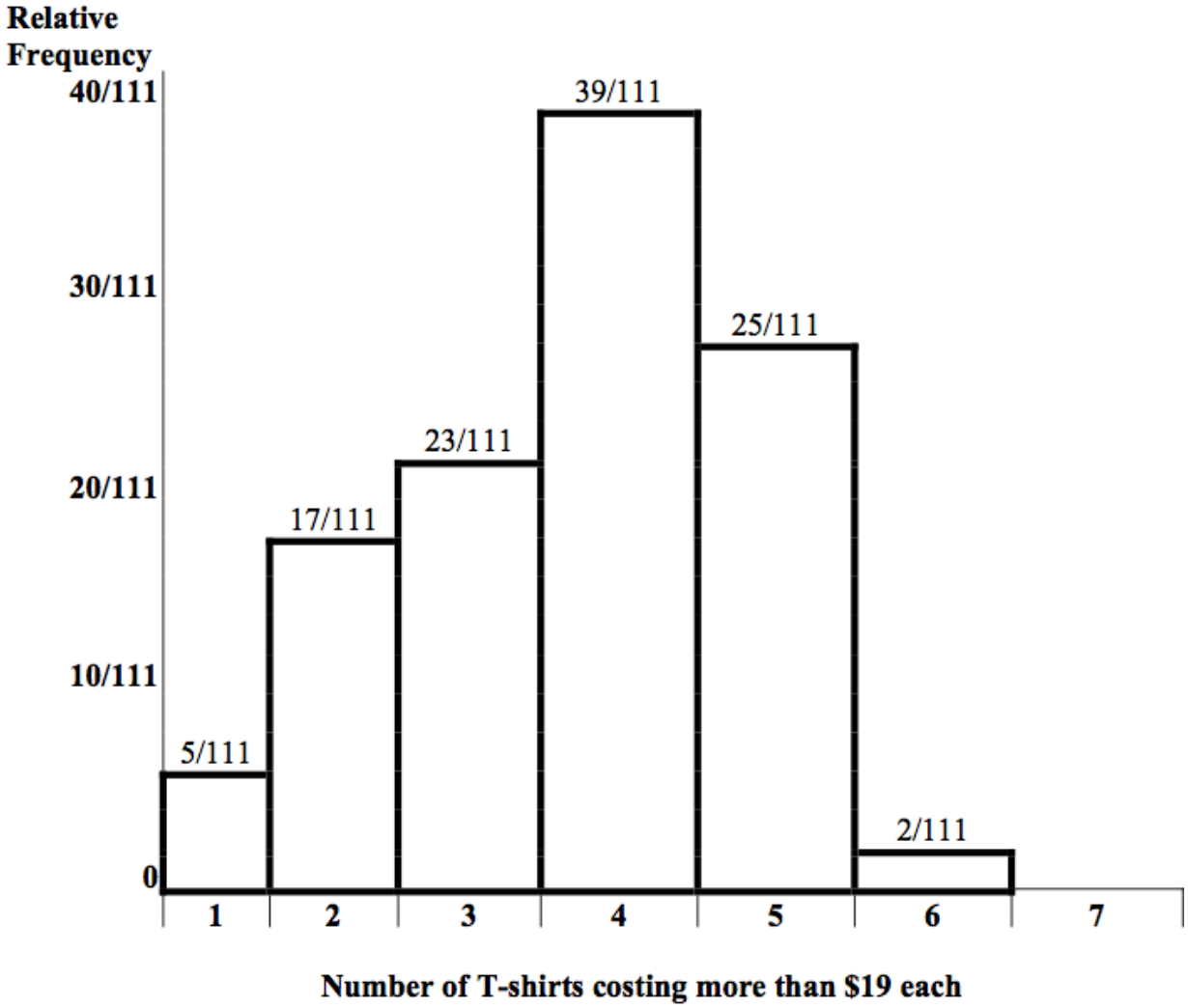
- **A**0.7

- **B**4.8
- **C**-2.8
- **D**Cannot be determined

Solution:

A

The next two questions refer to the following histogram. Suppose one hundred eleven people who shopped in a special T-shirt store were asked the number of T-shirts they own costing more than $19 each.



Exercise:

Problem:

The percent of people that own at most three (3) T-shirts costing more than $19 each is approximately:

- **A**21
- **B**59
- **C**41
- **D**Cannot be determined

Solution:

C

Exercise:

Problem:

If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

- **A**cluster
- **B**simple random
- **C**stratified
- **D**convenience

Solution:

D

Exercise:

Problem:

Below are the **2010 obesity rates by U.S. states and Washington, DC.**(*Source: http://www.cdc.gov/obesity/data/adult.html)*)

State	Percent (%)	State	Percent (%)
Alabama	32.2	Montana	23.0
Alaska	24.5	Nebraska	26.9
Arizona	24.3	Nevada	22.4
Arkansas	30.1	New Hampshire	25.0
California	24.0	New Jersey	23.8
Colorado	21.0	New Mexico	25.1
Connecticut	22.5	New York	23.9
Delaware	28.0	North Carolina	27.8
Washington, DC	22.2	North Dakota	27.2
Florida	26.6	Ohio	29.2
Georgia	29.6	Oklahoma	30.4
Hawaii	22.7	Oregon	26.8
Idaho	26.5	Pennsylvania	28.6
Illinois	28.2	Rhode Island	25.5
Indiana	29.6	South Carolina	31.5

State	Percent (%)	State	Percent (%)
Iowa	28.4	South Dakota	27.3
Kansas	29.4	Tennessee	30.8
Kentucky	31.3	Texas	31.0
Louisiana	31.0	Utah	22.5
Maine	26.8	Vermont	23.2
Maryland	27.1	Virginia	26.0
Massachusetts	23.0	Washington	25.5
Michigan	30.9	West Virginia	32.5
Minnesota	24.8	Wisconsin	26.3
Mississippi	34.0	Wyoming	25.1
Missouri	30.5		

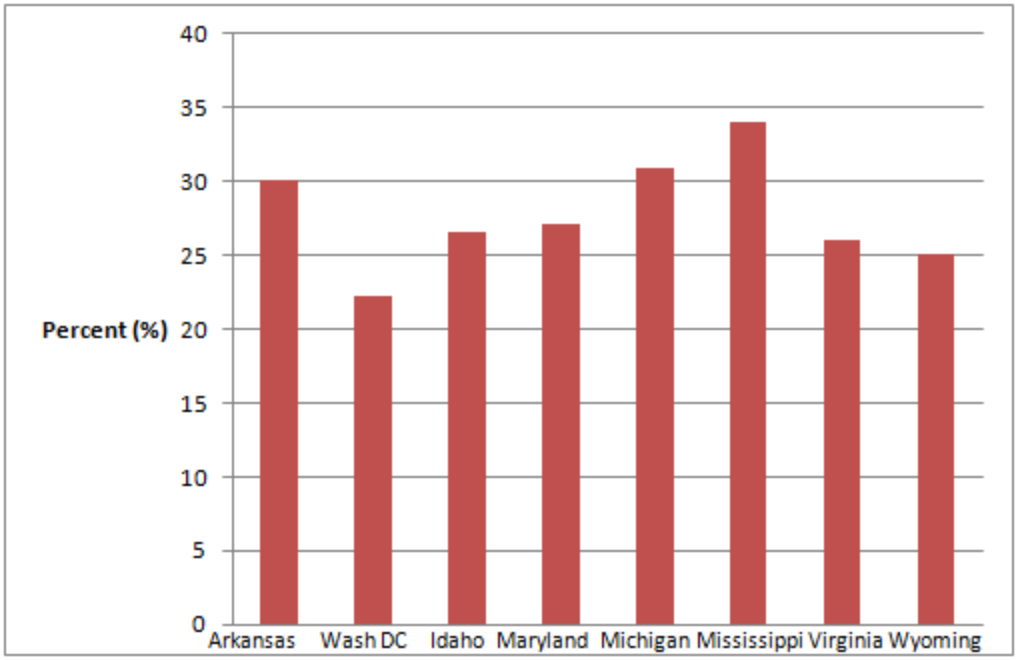
- **a.**Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the x-axis with the states.
- **b.**Use a random number generator to randomly pick 8 states. Construct a bar graph of the obesity rates of those 8 states.
- **c.**Construct a bar graph for all the states beginning with the letter "A."
- **d.**Construct a bar graph for all the states beginning with the letter "M."

Solution:

Example solution for **b** using the random number generator for the Ti-84 Plus to generate a simple random sample of 8 states. Instructions are below.

- Number the entries in the table 1 - 51 (Includes Washington, DC; Numbered vertically)
- Press MATH
- Arrow over to PRB
- Press 5:randInt(
- Enter 51,1,8)

Eight numbers are generated (use the right arrow key to scroll through the numbers). The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}. If any numbers are repeated, generate a different number by using 5:randInt(51,1)). Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}. Corresponding percents are {28.7 21.8 24.5 26 28.9 32.8 25 24.6}.



Exercise:

Problem:

A music school has budgeted to purchase 3 musical instruments. They plan to purchase a piano costing $3000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer numerically.

Solution:

For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar cost the most in comparison to the cost of other instruments of the same type.

Exercise:

Problem:

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the table below. (Note that this is the data presented for publisher B in homework exercise 13).

# of books	Freq.	Rel. Freq.

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Publisher B

a. Are there any outliers in the data? Use an appropriate numerical test involving the IQR to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than 2 standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts (a) and (c) of this problem give the same answer?
e. Examine the shape of the data. Which part, (a) or (c), of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

Solution:

- IQR = 4 – 1 = 3 ; Q1 – 1.5*IQR = 1 – 1.5(3) = -3.5 ; Q3 + 1.5*IQR = 4 + 1.5(3) = 8.5 ;The data value of 9 is larger than 8.5. The purchase of 9 books in one month is an outlier.
- The outlier should be investigated to see if there is an error or some other problem in the data; then a decision whether to include or exclude it should be made based on the particular situation. If it was a correct value then the data value should remain in the data set. If there is a problem with this data value, then it should be corrected or removed from the data. For example: If the data was recorded incorrectly (perhaps a 9 was miscoded and the correct value was 6) then the data should be corrected. If it was an error but the correct value is not known it should be removed from the data set.
- xbar – 2s = 2.45 – 2*1.88 = -1.31 ; xbar + 2s = 2.45 + 2*1.88 = 6.21 ; Using this method, the five data values of 7 books purchased and the one data value of 9 books purchased would be considered unusual.
- No: part (a) identifies only the value of 9 to be an outlier but part (c) identifies both 7 and 9.
- The data is skewed (to the right). It would be more appropriate to use the method involving the IQR in part (a), identifying only the one value of 9 books purchased as an outlier. Note that part (c) remarks that identifying unusual data values by using the criteria of being further than 2 standard deviations away from the mean is most appropriate when the data are mound-shaped and symmetric.
- The data are skewed to the right. For skewed data it is more appropriate to use the median as a measure of center.

**Exercises 32 and 33 contributed by Roberta Bloom

Lab: Descriptive Statistics

Class Time:

Names:

Student Learning Outcomes

- The student will construct a histogram and a box plot.
- The student will calculate univariate statistics.
- The student will examine the graphs to interpret what the data implies.

Collect the Data

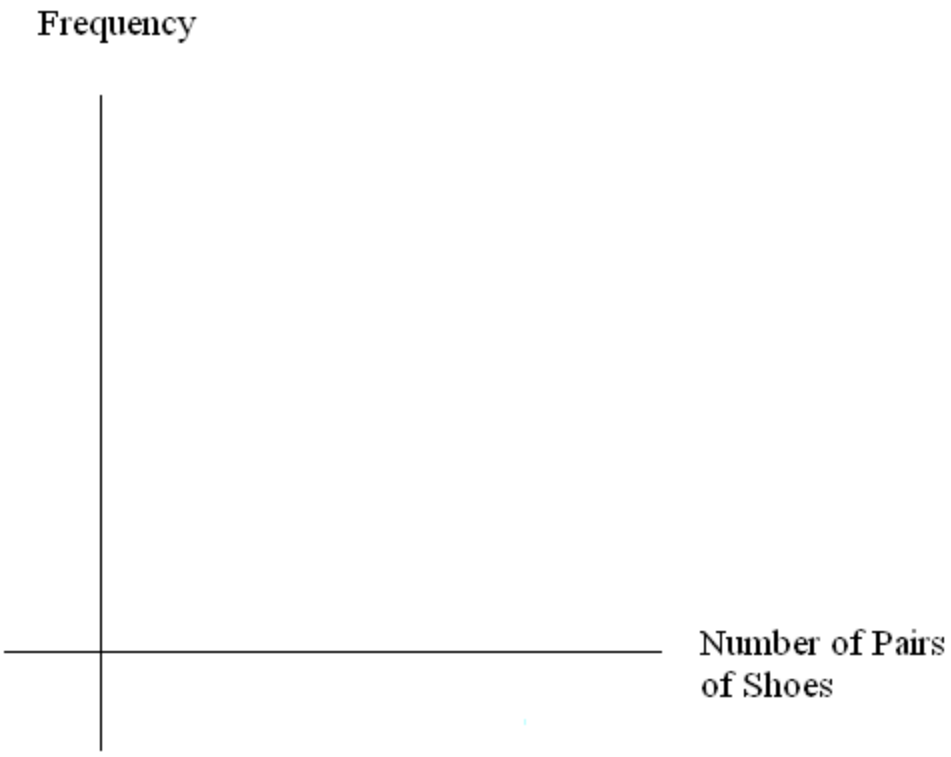
Record the number of pairs of shoes you own:

1. Randomly survey 30 classmates. Record their values.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Survey Results

2. Construct a histogram. Make 5-6 intervals. Sketch the graph using a ruler and pencil. Scale the axes.



3. Calculate the following:
    - x =
    - s =
4. Are the data discrete or continuous? How do you know?
5. Describe the shape of the histogram. Use complete sentences.
6. Are there any potential outliers? Which value(s) is (are) it (they)? Use a formula to check the end values to determine if they are potential outliers.

Analyze the Data

1. Determine the following:
    - Minimum value =

- Median =
- Maximum value =
- First quartile =
- Third quartile =
- IQR =

2. Construct a box plot of data
3. What does the shape of the box plot imply about the concentration of data? Use complete sentences.
4. Using the box plot, how can you determine if there are potential outliers?
5. How does the standard deviation help you to determine concentration of the data and whether or not there are potential outliers?
6. What does the IQR represent in this problem?
7. Show your work to find the value that is 1.5 standard deviations:

    - **a**Above the mean:
    - **b**Below the mean:

Linear Regression and Correlation
This module provides an introduction of Linear Regression and Correlation as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

Introduction

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is it and how strong is the relationship?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee. These are all examples in which regression can be used.

The type of data described in the examples is **bivariate** data - "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

Linear Regression and Correlation: Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. It has the form:
Equation:

$$y = a + \mathrm{bx}$$

where a and b are constant numbers.

x is the independent variable, and y is the dependent variable.
Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example:
The following examples are linear equations.
Equation:

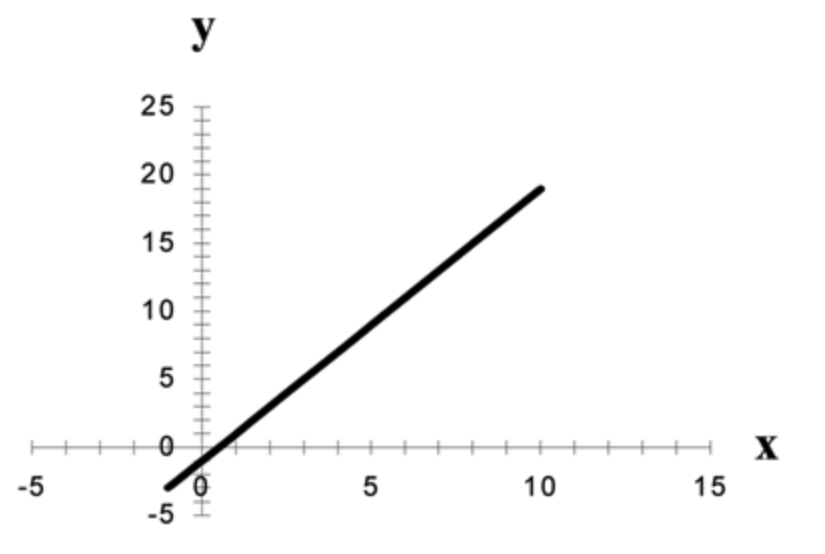
$$y = 3 + 2\mathrm{x}$$

Equation:

$$y = -0.01 + 1.2\mathrm{x}$$

The graph of a linear equation of the form $y = a + \mathrm{bx}$ is a **straight line**. Any line that is not vertical can be described by this equation.

Example:



Graph of the equation $y = -1 + 2x$.

Linear equations of this form occur in applications of life sciences, social sciences, psychology, business, economics, physical sciences, mathematics, and other areas.

Example:
Aaron's Word Processing Service (AWPS) does word processing. Its rate is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to do the word processing job.
Exercise:

Problem:

Find the equation that expresses the **total cost** in terms of the **number of hours** required to finish the word processing job.

Solution:

Let x = the number of hours it takes to get the job done.

Let y = the total cost to the customer.

The $31.50 is a fixed cost. If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only. The total cost is:

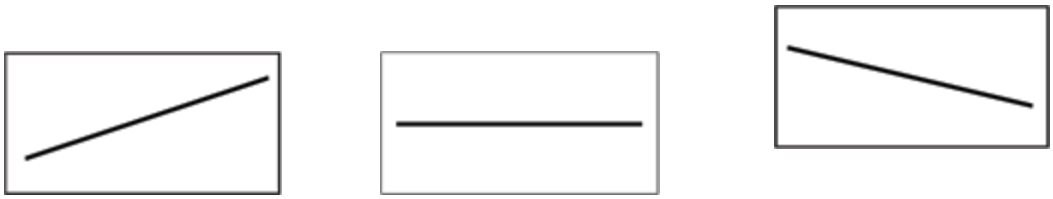
$y = 31.50 + 32\text{x}$

Linear Regression and Correlation: Slope and Y-Intercept of a Linear Equation

For the linear equation $y = a + \text{bx}$, b = slope and a = y-intercept.

From algebra recall that the slope is a number that describes the steepness of a line and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis.

If $b > 0$, the line slopes upward to the right.	If $b = 0$, the line is horizontal.	If $b < 0$, the line slopes downward to the right.



Three possible graphs of $y = a + \text{bx}$.

Example:
Svetlana tutors to make extra money for college. For each tutoring session, she charges a one time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15\text{x}$.

Exercise:

Problem:

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 (a = 25). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when x = 0). The slope is 15 (b = 15). For each session, Svetlana earns $15 for each hour she tutors.

Scatter Plots

This module provides an overview of Linear Regression and Correlation: Scatter Plots as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y. The most common and easiest way is a **scatter plot**. The following example illustrates a scatter plot.

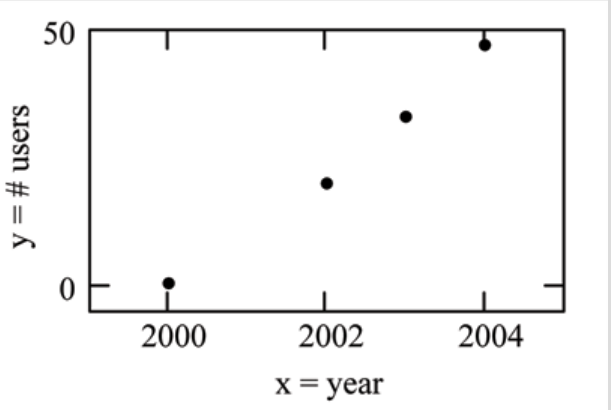
Example:

From an article in the *Wall Street Journal*: In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let x = the year and let y = the number of m-commerce users, in millions.

Table showing the number of m-commerce users (in millions) by year.

x (year)	y (# of users)
2000	0.5
2002	20.0

Scatter plot showing the number of m-commerce users (in millions) by year.



x (year)	y (# of users)
2003	33.0
2004	47.0

A scatter plot shows the **direction** and **strength** of a relationship between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the strength of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function.

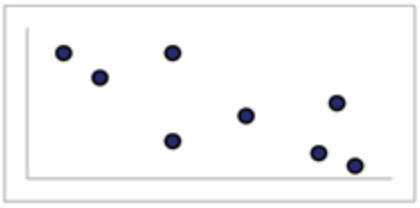
When you look at a scatterplot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.

Positive Linear Pattern (Strong)

Linear Pattern w/ One Deviation

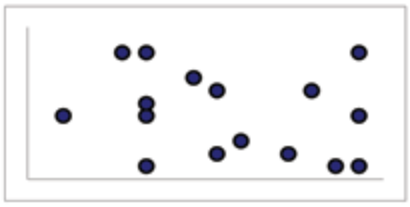
Negative Linear Pattern (Strong)

Negative Linear Pattern (Weak)



Exponential Growth Pattern

No Pattern



In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict

the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x.

The Regression Equation
Linear Regression and Correlation: The Regression Equation is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean. Contributions from Roberta Bloom include instructions for finding and graphing the regression equation and scatterplot using the LinRegTTest on the TI-83,83+,84+ calculators.

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This is called a **Line of Best Fit or Least Squares Line**.

Optional Collaborative Classroom Activity

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height.

For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your lines so they cross the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit". Do you think everyone will have the same equation? Why or why not?

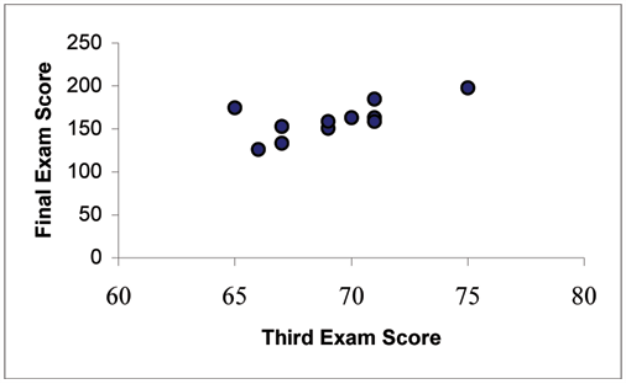
Using your equation, what is the predicted height for a pinky length of 2.5 inches?

Example:
A random sample of 11 statistics students produced the following data where x is the third exam score, out of 80, and y is the final exam score, out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Table showing the scores on the final exam based on scores from the third exam.

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

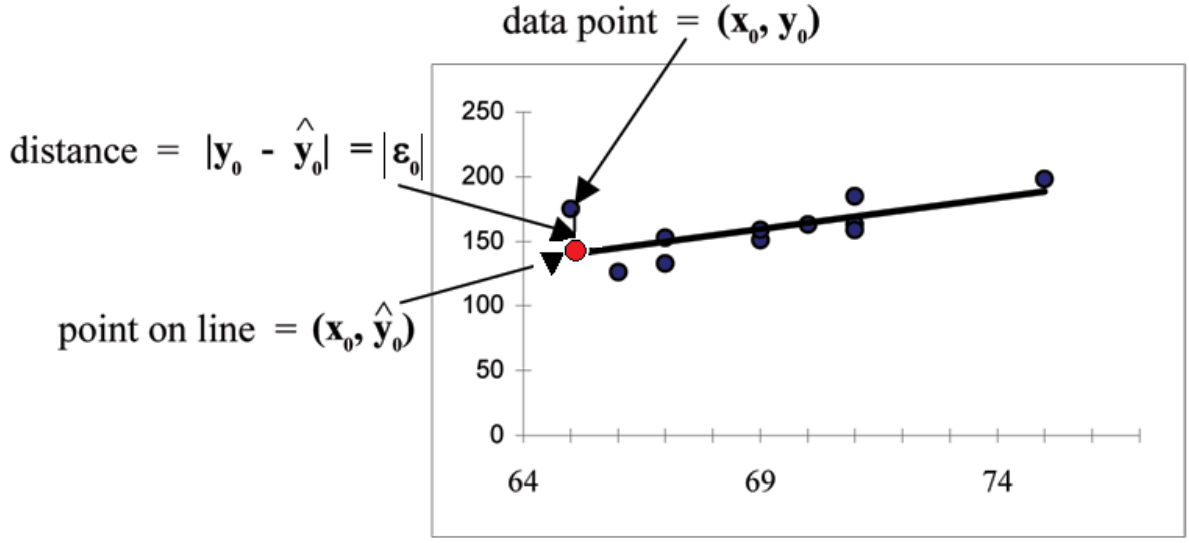
Scatter plot showing the scores on the final exam based on scores from the third exam.



The third exam score, x, is the independent variable and the final exam score, y, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye", you would draw different lines. We can use what is called a **least-squares regression line** to obtain the best fit line.

Consider the following diagram. Each point of data is of the the form (x, y) and each point of the line of best fit using least-squares linear regression has the form $(x, \hat{y})$.

The $\hat{y}$ is read **"y hat"** and is the **estimated value of** y. It is the value of y obtained using the regression line. It is not generally equal to y from data.



The term $y_0 - \hat{y}_0 = \varepsilon_0$ is called the **"error" or residual**. It is not an error in the sense of a mistake. The **absolute value of a residual** measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the diagram above, $y_0 - \hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

ε = the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1, 2, 3, ..., 11$.

Each $|\varepsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If you square each ε and add, you get

$$\left(\varepsilon_1\right)^2 + \left(\varepsilon_2\right)^2 + ... + \left(\varepsilon_{11}\right)^2 = \sum_{\mathrm{i}=1}^{11} \varepsilon^2$$

This is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:
Equation:

$$\hat{y} = a + \mathrm{bx}$$

where $a = \bar{y} - b \cdot \bar{x}$ and $b = \frac{\Sigma(x-\bar{x})\cdot(y-\bar{y})}{\Sigma(x-\bar{x})^2}$.

$\bar{x}$ and $\bar{y}$ are the sample means of the x values and the y values, respectively. The best fit line always passes through the point $(\bar{x}, \bar{y})$.

The slope b can be written as $b = r \cdot \left(\frac{s_y}{s_x}\right)$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the correlation coefficient which is discussed in the next section.

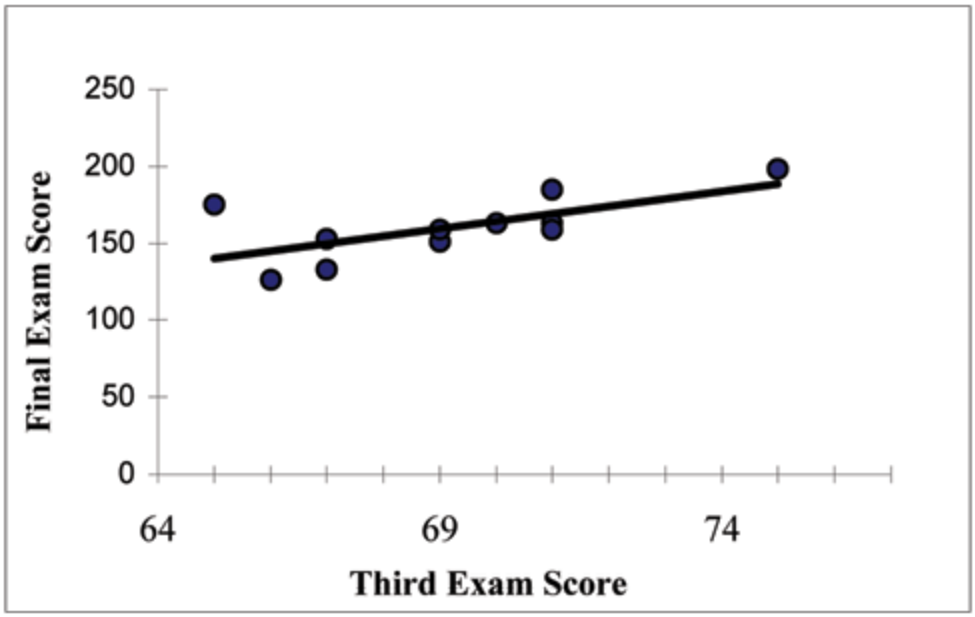
Least Squares Criteria for Best Fit

The process of fitting the best fit line is called **linear regression**. The idea behind finding the best fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least squares regression line** .

Note: Computer spreadsheets, statistical software, and many calculators can quickly calculate the best fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:

The graph of the line of best fit for the third exam/final exam example is shown below:



The least squares regression line (best fit line) for the third exam/final exam example has the equation:
Equation:

$$\hat{y} = -173.51 + 4.83\text{x}$$

Note:

- Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x-values in the sample data, **but not necessarily for x-values outside that domain.**
- You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam.
- You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x-values in the sample data, which are between 65 and 75.

UNDERSTANDING SLOPE
The slope of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.
THIRD EXAM vs FINAL EXAM EXAMPLE

- Slope: The slope of the line is b = 4.83.

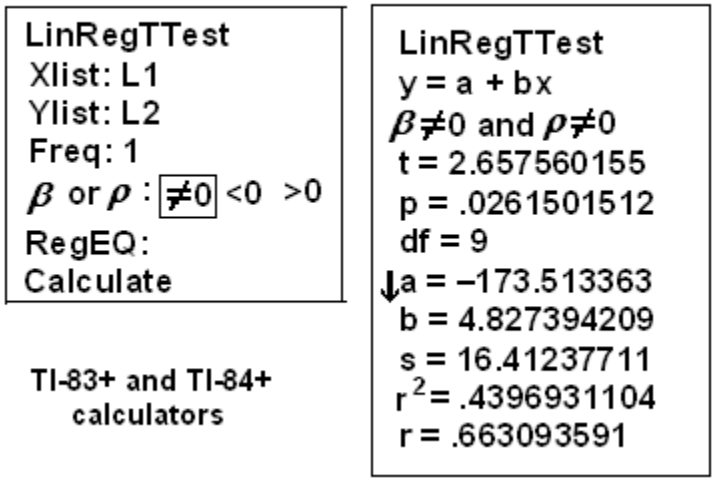
- Interpretation: For a one point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

Using the TI-83+ and TI-84+ Calculators

Using the Linear Regression T Test: LinRegTTest

In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x,y) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest as some calculators may also have a different item called LinRegTInt.)
On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1
On the next line, at the prompt β or ρ, highlight "≠ 0" and press ENTER
Leave the line for "RegEq:" blank
Highlight Calculate and press ENTER.

LinRegTTest Input Screen and Output Screen



TI-83+ and TI-84+ calculators

The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

- The second line says y=a+bx. Scroll down to find the values a=-173.513, and b=4.8273 ; the equation of the best fit line is $\hat{y} = -173.51 + 4.83x$
- The two items at the bottom are r^2 = .43969 and r=.663. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

We are assuming your X data is already entered in list L1 and your Y data is in list L2
Press 2nd STATPLOT ENTER to use Plot 1
On the input screen for PLOT 1, highlight**On**and press ENTER
For TYPE: highlight the very first icon which is the scatterplot and press ENTER
Indicate Xlist: L1 and Ylist: L2
For Mark: it does not matter which symbol you highlight.
Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data
To graph the best fit line, press the "Y=" key and type the equation -173.5+4.83X into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.
Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

**With contributions from Roberta Bloom

Correlation Coefficient and Coefficient of Determination
Linear Regression and Correlation: The Correlation Coefficient and Coefficient of Determination is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom. The name has been changed from Correlation Coefficient.

The Correlation Coefficient r

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y.

The **correlation coefficient, r,** developed by Karl Pearson in the early 1900s, is a numerical measure of the strength of association between the independent variable x and the dependent variable y.

The correlation coefficient is calculated as
Equation:

$$r = \frac{n \cdot \Sigma x \cdot y - (\Sigma x) \cdot (\Sigma y)}{\sqrt{[n \cdot \Sigma x^2 - (\Sigma x)^2] \cdot [n \cdot \Sigma y^2 - (\Sigma y)^2]}}$$

where n = the number of data points.

If you suspect a linear relationship between x and y, then r can measure how strong the linear relationship is.
What the VALUE of r tells us:

- The value of r is always between -1 and +1: $-1 \leq r \leq 1$.
- The size of the correlation r indicates the strength of the linear relationship between x and y. Values of r close to -1 or to +1 indicate a stronger linear relationship between x and y.
- If $r=0$ there is absolutely no linear relationship between x and y **(no linear correlation)**.

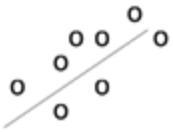
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

What the SIGN of r tells us

- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease **(positive correlation)**.
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase **(negative correlation)**.
- The sign of r is the same as the sign of the slope, b, of the best fit line.

Note: Strong correlation does not suggest that x causes y or y causes x. We say **"correlation does not imply causation."** For example, every person who learned math in the 17th century is dead. However, learning math does not necessarily cause death!

Positive Correlation

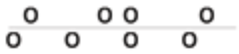


A scatter plot showing data with a positive correlation.
$0 < r < 1$

Negative Correlation

A scatter plot showing data with a negative correlation. $-1 < r < 0$

Zero Correlation



A scatter plot showing data with zero correlation. r =0

The formula for r looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate r. The correlation coefficient r is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

r^2 is called the coefficient of determination. r^2 is the square of the correlation coefficient , but is usually stated as a percent, rather than in decimal form. r^2 has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression (best fit) line.
- 1-r^2, when expressed as a percent, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the [third exam/final exam example] introduced in the previous section

- The line of best fit is: $\hat{y} = -173.51 + 4.83\text{x}$
- The correlation coefficient is $r = 0.6631$
- The coefficient of determination is r^2 = 0.6631^2 = 0.4397
- **Interpretation of r^2 in the context of this example:**
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final exam grades can be explained by the variation in the grades on the third exam, using the best fit regression line.
- Therefore approximately 56% of the variation (1 - 0.44 = 0.56) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best fit regression line. (This is seen as the scattering of the points about the line.)

**With contributions from Roberta Bloom.

Glossary

Coefficient of Correlation
: A measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable. The formula is:
Equation:

$$r = \frac{n \sum \mathrm{xy} - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}},$$

where n is the number of data points. The coefficient cannot be more then 1 and less then -1. The closer the coefficient is to ± 1, the stronger the evidence of a significant linear relationship between x and y.

Testing the Significance of the Correlation Coefficient
Linear Regression and Correlation: Testing the Significance of the Correlation Coefficient is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean. The title has been changed from Facts About the Correlation Coefficient for Linear Regression. Roberta Bloom has made major contributions to this module.

Testing the Significance of the Correlation Coefficient

The correlation coefficient, r, tells us about the strength of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n, together.

We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data is used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we only have sample data, we can not calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

- The symbol for the population correlation coefficient is ρ, the Greek letter "rho".
- ρ = population correlation coefficient (unknown)
- r = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to 0" or "significantly different from 0". We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes that the correlation coefficient is significantly different from 0, we say that the correlation coefficient is "significant".

- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0."
- What the conclusion means: There is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from 0 (it is close to 0), we say that correlation coefficient is "not significant".

- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from 0."
- What the conclusion means: There is not a significant linear relationship between x and y. Therefore we can NOT use the regression line to model a linear relationship between x and y in the population.

Note:

- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant OR if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If r is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

PERFORMING THE HYPOTHESIS TEST
SETTING UP THE HYPOTHESES:

- **Null Hypothesis:** H_o**:** ρ **= 0**
- **Alternate Hypothesis:** H_a**:** $\rho \neq$ **0**

What the hypotheses mean in words:

- **Null Hypothesis** H_o**:** The population correlation coefficient IS NOT significantly different from 0. There IS NOT a significant linear relationship(correlation) between x and y in the population.
- **Alternate Hypothesis** H_a**:** The population correlation coefficient IS significantly DIFFERENT FROM 0. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

DRAWING A CONCLUSION:

- There are two methods to make the decision. Both methods are equivalent and give the same result.
- **Method 1: Using the p-value**
- **Method 2: Using a table of critical values**
- In this chapter of this textbook, we will always use a significance level of 5%, α = 0.05
- Note: Using the p-value method, you could choose any appropriate significance level you want; you are not limited to using α = 0.05. But the table of critical values provided in this textbook assumes that we are using a significance level of 5%, α = 0.05. (If we wanted to use a different significance level than 5% with the critical value method, we would need different tables of critical values that are not provided in this textbook.)

METHOD 1: Using a p-value to make a decision

- The linear regression t-test LinRegTTEST on the TI-83+ or TI-84+ calculators calculates the p-value.
- On the LinRegTTEST input screen, on the line prompt for β or ρ, highlight "$\neq$ **0**"

- The output screen shows the p-value on the line that reads "p =".
- (Most computer statistical software can calculate the p-value.)

If the p-value is less than the significance level (α = 0.05):

- Decision: REJECT the null hypothesis.
- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0."

If the p-value is NOT less than the significance level (α = 0.05)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from 0."

Calculation Notes:

- You will use technology to calculate the p-value. The following describe the calculations to compute the test statistics and the p-value:
- The p-value is calculated using a t-distribution with n-2 degrees of freedom.
- The formula for the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, t, is shown in the computer or calculator output along with the p-value. The test statistic t has the same sign as the correlation coefficient r.
- The p-value is the combined area in both tails.
- An alternative way to calculate the p-value **(p)** given by LinRegTTest is the command 2*tcdf(abs(t),10^99, n-2) in 2nd DISTR.

THIRD EXAM vs FINAL EXAM EXAMPLE: p value method

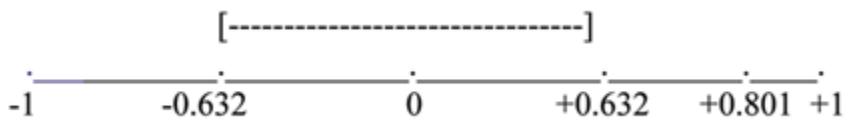
- Consider the third exam/final exam example.
- The line of best fit is: $\hat{y} = -173.51 + 4.83$x with $r = 0.6631$ and there are $n = 11$ data points.

- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

- H_o: $\rho = 0$
- H_a: $\rho \neq 0$
- $\alpha = 0.05$
- The p-value is 0.026 (from LinRegTTest on your calculator or from computer software)
- The p-value, 0.026, is less than the significance level of $\alpha = 0.05$
- Decision: Reject the Null Hypothesis H_o
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0.
- **Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.**

METHOD 2: Using a table of Critical Values to make a decision
The 95% Critical Values of the Sample Correlation Coefficient Table at the end of this chapter (before the Summary) may be used to give you a good idea of whether the computed value of r **is significant or not**. Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may want to use the line for prediction.

Example:
Suppose you computed $r = 0.801$ using $n = 10$ data points.
df $= n - 2 = 10 - 2 = 8$. The critical values associated with df $= 8$ are -0.632 and + 0.632. If $r <$ negative critical value or $r >$ positive critical value, then r is significant. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.

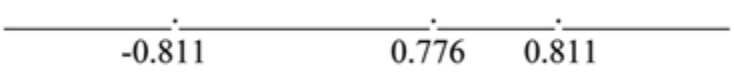


r is not significant between -0.632 and +0.632.
$r = 0.801 > +0.632$. Therefore, r is significant.

Example:
Suppose you computed $r = -0.624$ with 14 data points.
$\text{df} = 14 - 2 = 12$. The critical values are -0.532 and 0.532. Since $-0.624 < -0.532$, r is significant and the line may be used for prediction

$r = -0.624 < -0.532$. Therefore, r is significant.

Example:
Suppose you computed $r = 0.776$ and $n = 6$. $\text{df} = 6 - 2 = 4$. The critical values are -0.811 and 0.811. Since $-0.811 < 0.776 < 0.811$, r is not significant and the line should not be used for prediction.



$-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

THIRD EXAM vs FINAL EXAM EXAMPLE: critical value method

- Consider the third exam/final exam example.
- The line of best fit is: $\hat{y} = -173.51 + 4.83\text{x}$ with $r = 0.6631$ and there are $\text{n} = 11$ data points.
- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

- H_o: $\rho = 0$
- H_a: $\rho \neq 0$
- $\alpha = 0.05$
- Use the "95% Critical Value" table for r with $\text{df} = \text{n} - 2 = 11 - 2 = 9$
- The critical values are -0.602 and +0.602
- Since $0.6631 > 0.602$, r is significant.
- Decision: Reject H_o:
- Conclusion:There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0.
- **Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.**

Example:
Additional Practice Examples using Critical Values
Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

1. $r = -0.567$ and the sample size, n, is 19. The $\text{df} = n - 2 = 17$. The critical value is -0.456. $-0.567 < -0.456$ so r is significant.
2. $r = 0.708$ and the sample size, n, is 9. The $\text{df} = n - 2 = 7$. The critical value is 0.666. $0.708 > 0.666$ so r is significant.
3. $r = 0.134$ and the sample size, n, is 14. The $\text{df} = 14 - 2 = 12$. The critical value is 0.532. 0.134 is between -0.532 and 0.532 so r is not

significant.
4. $r = 0$ and the sample size, n, is 5. No matter what the dfs are, $r = 0$ is between the two critical values so r is not significant.

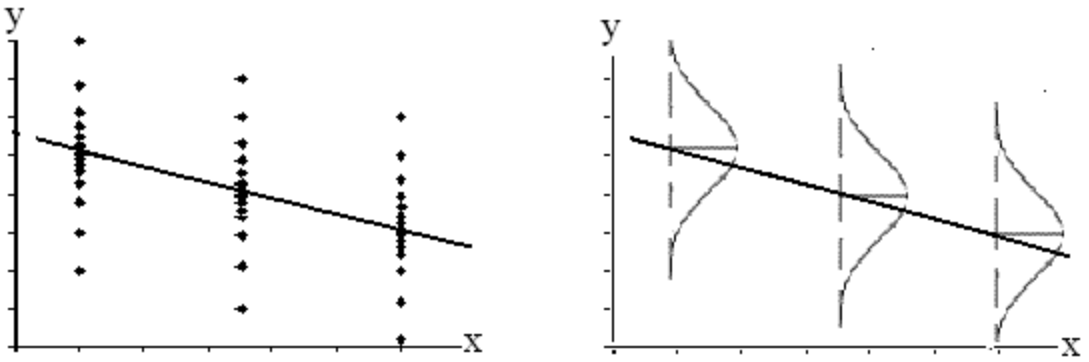
Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

The regression line equation that we calculate from the sample data gives the best fit line for our particular sample. We want to use this best fit line for the sample as an estimate of the best fit line for the population. Examining the scatterplot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.
The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x. In other words, the expected value of y for each particular value lies on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) above implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.

- The standard deviations of the population y values about the line are equal for each value of x. In other words, each of these normal distributions of y values has the same shape and spread about the line.
- The residual errors are mutually independent (no pattern).



The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.

**With contributions from Roberta Bloom

Prediction
Linear Regression and Correlation: Prediction is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom.

Recall the third exam/final exam example.

We examined the scatterplot and showed that the correlation coefficient is significant. We found the equation of the best fit line for the final exam grade as a function of the grade on the third exam. We can now use the least squares regression line for prediction.

Suppose you want to estimate, or predict, the final exam score of statistics students who received 73 on the third exam. The exam scores **(x-values)** range from 65 to 75. **Since 73 is between the x-values 65 and 75**, substitute $x = 73$ into the equation. Then:
Equation:

$$\hat{y} = -173.51 + 4.83(73) = 179.08$$

We predict that statistic students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

Example:
Recall the third exam/final exam example.
Exercise:

Problem:

What would you predict the final exam score to be for a student who scored a 66 on the third exam?

Solution:

145.27

Exercise:

Problem:

What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Solution:

The x values in the data are between 65 and 75. 90 is outside of the domain of the observed x values in the data (independent variable), so you cannot reliably predict the final exam score for this student. (Even though it is possible to enter x into the equation and calculate a y value, you should not do so!)

To really understand how unreliable the prediction can be outside of the observed x values in the data, make the substitution x = 90 into the equation.

$$\hat{y} = -173.51 + 4.83\big(90\big) = 261.19$$

The final exam score is predicted to be 261.19. The largest the final exam score can be is 200.

Note: The process of predicting inside of the observed x values in the data is called **interpolation**. The process of predicting outside of the observed x values in the data is called **extrapolation**.

**With contributions from Roberta Bloom

Outliers

Linear Regression and Correlation: Outliers is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean. The module has been modified to include a graphical method for identifying outliers contributed by Roberta Bloom.

In some data sets, there are values **(observed data points)** called **outliers**. **Outliers are observed data points that are far from the least squares line.** They have large "errors", where the "error" or residual is the vertical distance from the line to the point.

Outliers need to be examined closely. Sometimes, for some reason or another, they should not be included in the analysis of the data. It is possible that an outlier is a result of erroneous data. Other times, an outlier may hold valuable information about the population under study and should remain included in the data. The key is to carefully examine what causes a data point to be an outlier.

Besides outliers, a sample may contain one or a few points that are called **influential points**. Influential points are observed data points that are far from the other observed data points in the horizontal direction. These points may have a big effect on the slope of the regression line. To begin to identify an influential point, you can remove it from the data set and see if the slope of the regression line is changed significantly.

Computers and many calculators can be used to identify outliers from the data. Computer output for regression analysis will often identify both outliers and influential points so that you can examine them.

Identifying Outliers

We could guess at outliers by looking at a graph of the scatterplot and best fit line. However we would like some guideline as to how far away a point needs to be in order to be considered an outlier. **As a rough rule of thumb, we can flag any point that is located further than two standard deviations above or below the best fit line as an outlier**. The standard deviation used is the standard deviation of the residuals or errors.

We can do this visually in the scatterplot by drawing an extra pair of lines that are two standard deviations above and below the best fit line. Any data points that are outside this extra pair of lines are flagged as potential outliers. Or we can do this numerically by calculating each residual and comparing it to twice the standard deviation. On the TI-83, 83+, or 84+, the graphical approach is easier. The graphical procedure is shown first, followed by the numerical calculations. You would generally only need to use one of these methods.

Example:
Exercise:

Problem:

In the third exam/final exam example, you can determine if there is an outlier or not. If there is an outlier, as an exercise, delete it and fit the remaining data to a new line. For this example, the new line ought to fit the remaining data better. This means the **SSE** should be smaller and the correlation coefficient ought to be closer to 1 or -1.

Solution:

Graphical Identification of Outliers
With the TI-83,83+,84+ graphing calculators, it is easy to identify the outlier graphically and visually. If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance was equal to $2s$ or farther, then we would consider the data point to be "too far" from the line of best fit. We need to find and graph the lines that are two standard deviations below and above the regression line. Any points that are outside these two lines are outliers. We will call these lines Y2 and Y3:

As we did with the equation of the regression line and the correlation coefficient, we will use technology to calculate this standard deviation for us. Using the **LinRegTTest** with this data, scroll down through the output screens to find **s=16.412**

Line Y2=-173.5+4.83x-2(16.4) and line Y3=-173.5+4.83x+2(16.4)

where $\hat{y}$=-173.5+4.83x is the line of best fit. Y2 and Y3 have the same slope as the line of best fit.

Graph the scatterplot with the best fit line in equation Y1, then enter the two extra lines as Y2 and Y3 in the "Y="equation editor and press ZOOM 9. You will find that the only data point that is not between lines Y2 and Y3 is the point x=65, y=175. On the calculator screen it is just barely outside these lines. The outlier is the student who had a grade of 65 on the third exam and 175 on the final exam; this point is further than 2 standard deviations away from the best fit line.

Sometimes a point is so close to the lines used to flag outliers on the graph that it is difficult to tell if the point is between or outside the lines. On a computer, enlarging the graph may help; on a small calculator screen, zooming in may make the graph clearer. Note that when the graph does not give a clear enough picture, you can use the numerical comparisons to identify outliers.
[missing_resource: linrgoutlier.gif]

Numerical Identification of Outliers
In the table below, the first two columns are the third exam and final exam data. The third column shows the predicted $\hat{y}$ values calculated from the line of best fit: $\hat{y}$=-173.5+4.83x. The residuals, or errors, have been calculated in the fourth column of the table:
$\text{observed y value} - \text{predicted y value} = y - \hat{y}$.

s is the standard deviation of all the $y - \hat{y} = \varepsilon$ values where n = the total number of data points. If each residual is calculated and squared, and the results are added, we get the SSE. The standard deviation of the residuals is calculated from the SSE as:

$$s = \sqrt{\frac{\text{SSE}}{n-2}}$$

Rather than calculate the value of s ourselves, we can find s using the computer or calculator. For this example, the calculator function

LinRegTTest found s = 16.4 as the standard deviation of the residuals 35 -17 16 -6 -19 9 3 -1 -10 -9 -1 .

x	y	$\hat{y}$	$y - \hat{y}$
65	175	140	$175 - 140 = 35$
67	133	150	$133 - 150 = -17$
71	185	169	$185 - 169 = 16$
71	163	169	$163 - 169 = -6$
66	126	145	$126 - 145 = -19$
75	198	189	$198 - 189 = 9$
67	153	150	$153 - 150 = 3$
70	163	164	$163 - 164 = -1$
71	159	169	$159 - 169 = -10$
69	151	160	$151 - 160 = -9$
69	159	160	$159 - 160 = -1$

We are looking for all data points for which the residual is greater than 2s=2(16.4)=32.8 or less than -32.8. Compare these values to the residuals in column 4 of the table. The only such data point is the student who had a grade of 65 on the third exam and 175 on the final exam; the residual for this student is 35.

How does the outlier affect the best fit line?
Numerically and graphically, we have identified the point (65,175) as an outlier. We should re-examine the data for this point to see if there are any problems with the data. If there is an error we should fix the error if possible, or delete the data. If the data is correct, we would leave it in the data set. **For this problem, we will suppose that we examined the data and found that this outlier data was an error. Therefore we will continue on and delete the outlier, so that we can explore how it affects the results, as a learning experience.**

Compute a new best-fit line and correlation coefficient using the 10 remaining points:
On the TI-83, TI-83+, TI-84+ calculators, delete the outlier from L1 and L2. Using the LinRegTTest, the new line of best fit and the correlation coefficient are:

$\hat{y} = -355.19 + 7.39$x and $r = 0.9121$

The new line with $r = 0.9121$ is a stronger correlation than the original (r=0.6631) because $r = 0.9121$ is closer to 1. This means that the new line is a better fit to the 10 remaining data values. The line can better predict the final exam score given the third exam score.

Numerical Identification of Outliers: Calculating s and Finding Outliers Manually

If you do not have the function LinRegTTest, then you can calculate the outlier in the first example by doing the following.

First, **square each** $|y - \hat{y}|$ (See the TABLE above):

The squares are 35^2 17^2 16^2 6^2 19^2 9^2 3^2 1^2 10^2 9^2 1^2

Then, add (sum) all the $|y - \hat{y}|$ **squared terms** using the formula

$\sum_{i=1}^{11} \left(\left| y_i - \hat{y}_i \right| \right)^2 = \sum_{i=1}^{11} {\varepsilon_i}^2$ (Recall that $y_i - \hat{y}_i = \varepsilon_i$.)

$= 35^2 + 17^2 + 16^2 + 6^2 + 19^2 + 9^2 + 3^2 + 1^2 + 10^2 + 9^2 + 1^2$

$= 2440 =$ **SSE**. The result, **SSE** is the Sum of Squared Errors.

Next, calculate s, the standard deviation of all the $y - \hat{y} = \varepsilon$ values where n = the total number of data points.

The calculation is $s = \sqrt{\frac{\text{SSE}}{n-2}}$

For the third exam/final exam problem, $s = \sqrt{\frac{2440}{11-2}} = 16.47$

Next, multiply s by 1.9:
$(1.9) \cdot (16.47) = 31.29$
31.29 is almost 2 standard deviations away from the mean of the $y - \hat{y}$ values.

If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance is at least $1.9s$, then we would consider the data point to be "too far" from the line of best fit. We call that point a **potential outlier**.

For the example, if any of the $\left| y - \hat{y} \right|$ values are **at least** 31.29, the corresponding (x, y) data point is a potential outlier.

For the third exam/final exam problem, all the $\left| y - \hat{y} \right|$'s are less than 31.29 except for the first one which is 35.

$35 > 31.29$ That is, $\left| y - \hat{y} \right| \geq \left(1.9\right) \cdot \left(s\right)$

The point which corresponds to $\left| y - \hat{y} \right| = 35$ is $(65, 175)$. **Therefore, the data point $(65, 175)$ is a potential outlier.** For this example, we will delete it. (Remember, we do not always delete an outlier.)

The next step is to compute a new best-fit line using the 10 remaining points. The new line of best fit and the correlation coefficient are:

$\hat{y} = -355.19 + 7.39x$ and $r = 0.9121$

Example:
Exercise:

Problem:

Using this new line of best fit (based on the remaining 10 data points), what would a student who receives a 73 on the third exam expect to receive on the final exam? Is this the same as the prediction made using the original line?

Solution:

Using the new line of best fit, $\hat{y} = -355.19 + 7.39(73) = 184.28$. A student who scored 73 points on the third exam would expect to earn 184 points on the final exam.

The original line predicted $\hat{y} = -173.51 + 4.83(73) = 179.08$ so the prediction using the new line with the outlier eliminated differs from the original prediction.

Example:
(*From The Consumer Price Indexes Web site*) The Consumer Price Index (CPI) measures the average change over time in the prices paid by urban consumers for consumer goods and services. The CPI affects nearly all Americans because of the many ways it is used. One of its biggest uses is as a measure of inflation. By providing information about price changes in the Nation's economy to government, business, and labor, the CPI helps them to make economic decisions. The President, Congress, and the

Federal Reserve Board use the CPI's trends to formulate monetary and fiscal policies. In the following table, x is the year and y is the CPI.

x	y
1915	10.1
1926	17.7
1935	13.7
1940	14.7
1947	24.1
1952	26.5
1964	31.0
1969	36.7
1975	49.3
1979	72.6
1980	82.4
1986	109.6
1991	130.7
1999	166.6

Data:

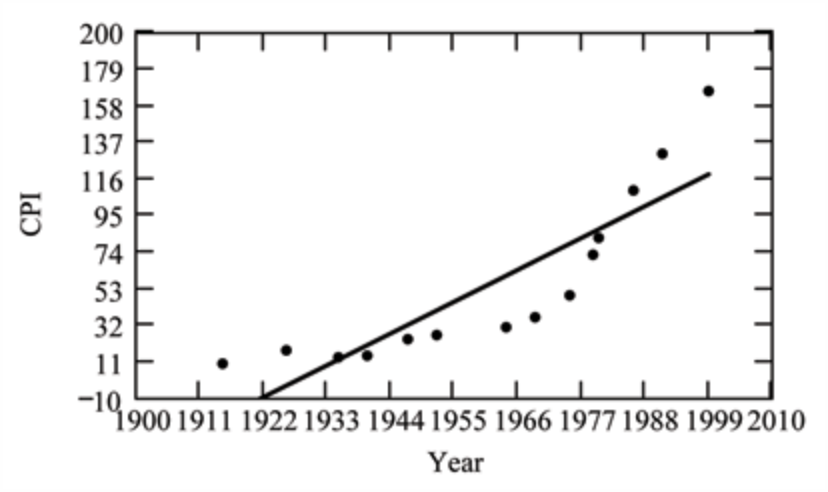
Exercise:

Problem:

- Make a scatterplot of the data.
- Calculate the least squares line. Write the equation in the form $\hat{y} = a + \mathrm{b}x$.
- Draw the line on the scatterplot.
- Find the correlation coefficient. Is it significant?
- What is the average CPI for the year 1990?

Solution:

- Scatter plot and line of best fit.
- $\hat{y} = -3204 + 1.662x$ is the equation of the line of best fit.
- $r = 0.8694$
- The number of data points is $n = 14$. Use the 95% Critical Values of the Sample Correlation Coefficient table at the end of Chapter 12. $n - 2 = 12$. The corresponding critical value is 0.532. Since $0.8694 > 0.532$, r is significant.
- $\hat{y} = -3204 + 1.662\big(1990\big) = 103.4$ CPI
- Using the calculator LinRegTTest, we find that s = 25.4 ; graphing the lines Y2=-3204+1.662X-2(25.4) and Y3=-3204+1.662X+2(25.4) shows that no data values are outside those lines, identifying no outliers. (Note that the year 1999 was very close to the upper line, but still inside it.)



Note: In the example, notice the pattern of the points compared to the line. Although the correlation coefficient is significant, the pattern in the scatterplot indicates that a curve would be a more appropriate model to use than a line. In this example, a statistician should prefer to use other methods to fit a curve to this data, rather than model the data with the line we found. In addition to doing the calculations, it is always important to look at the scatterplot when deciding whether a linear model is appropriate.

If you are interested in seeing more years of data, visit the Bureau of Labor Statistics CPI website ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt ; our data is taken from the column entitled "Annual Avg." (third column from the right). For example you could add more current years of data. Try adding the more recent years 2004 : CPI=188.9, 2008 : CPI=215.3 and 2011: CPI=224.9. See how it affects the model. (Check: $\hat{y} = -4436 + 2.295x$. r $= 0.9018$. Is r significant? Is the fit better with the addition of the new points?)

**With contributions from Roberta Bloom

Glossary

Outlier
: An observation that does not fit the rest of the data.

95% Critical Values of the Sample Correlation Coefficient Table
This module provides an overview of Linear Regression and Correlation: 95% Critical Values of the Sample Correlation Coefficient Table as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Degrees of Freedom:	**Critical Values: (and)**
1	0.997
2	0.950
3	0.878
4	0.811
5	0.754
6	0.707
7	0.666
8	0.632
9	0.602
10	0.576
11	0.555
12	0.532

Degrees of Freedom:	Critical Values: (and)
13	0.514
14	0.497
15	0.482
16	0.468
17	0.456
18	0.444
19	0.433
20	0.423
21	0.413
22	0.404
23	0.396
24	0.388
25	0.381
26	0.374
27	0.367
28	0.361
29	0.355

Degrees of Freedom:	**Critical Values: (and)**
30	0.349
40	0.304
50	0.273
60	0.250
70	0.232
80	0.217
90	0.205
100	0.195

Linear Regression and Correlation: Summary

Bivariate Data: Each data point has two values. The form is (x, y).

Line of Best Fit or Least Squares Line (LSL): $\hat{y} = a + \text{bx}$

x = independent variable; y = dependent variable

Residual: $\text{Actual y value} - \text{predicted y value} = y - \hat{y}$
Correlation Coefficient r:

1. Used to determine whether a line of best fit is good for prediction.
2. Between -1 and 1 inclusive. The closer r is to 1 or -1, the closer the original points are to a straight line.
3. If r is negative, the slope is negative. If r is positive, the slope is positive.
4. If $r = 0$, then the line is horizontal.

Sum of Squared Errors (SSE): The smaller the **SSE**, the better the original set of points fits the line of best fit.

Outlier: A point that does not seem to fit the rest of the data.

Practice: Linear Regression
This module provides a practice of Linear Regression and Correlation as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

- The student will evaluate bivariate data and determine if a line is an appropriate fit to the data.

Given

Below are real data for the first two decades of AIDS reporting. (*Source: Centers for Disease Control and Prevention, National Center for HIV, STD, and TB Prevention*)

Year	**# AIDS cases diagnosed**	**# AIDS deaths**
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987

1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Adults and Adolescents only, United States

Note: We will use the columns “year” and “# AIDS cases diagnosed” for all questions unless otherwise stated.

Graphing

Graph “year” vs. “# AIDS cases diagnosed.” **Plot the points on the graph located below in the section titled "Plot"** . Do not include pre-1981. Label both axes with words. Scale both axes.

Data

Exercise:

Problem:

Enter your data into your calculator or computer. The pre-1981 data should not be included. Why is that so?

Linear Equation

Write the linear equation below, rounding to 4 decimal places:

Note: For any prediction questions, the answers are calculated using the least squares (best fit) line equation cited in the solution.

Exercise:

Problem: Calculate the following:

- **a** $a =$
- **b** $b =$

- **c** corr. $=$
- **d** $n =$(# of pairs)

Solution:

- **a** $a =$ -3,448,225
- **b** $b = 1750$
- **c** corr. $= 0.4526$
- **d** $n = 22$

Exercise:

Problem: equation: $\hat{y} =$

Solution:

$\hat{y} =$ -3,448,225 $+1750x$

Solve

Exercise:

Problem: Solve.

- **a** When $x = 1985$, $\hat{y} =$
- **b** When $x = 1990$, $\hat{y} =$

Solution:

- **a** 25,525
- **b** 34,275

Plot

Plot the 2 above points on the graph below. Then, connect the 2 points to form the regression line.

Obtain the graph on your calculator or computer.

Discussion Questions

Look at the graph above.

Exercise:

Problem: Does the line seem to fit the data? Why or why not?

Exercise:

Problem: Do you think a linear fit is best? Why or why not?

Exercise:

Problem:

Hand draw a smooth curve on the graph above that shows the flow of the data.

Exercise:

Problem:

What does the correlation imply about the relationship between time (years) and the number of diagnosed AIDS cases reported in the U.S.?

Exercise:

Problem:

Why is “year” the independent variable and “# AIDS cases diagnosed.” the dependent variable (instead of the reverse)?

Exercise:

Problem: Solve.

- **a**When $x = 1970$, $\hat{y} =$:
- **b**Why doesn’t this answer make sense?

Solution:

- **a** -725

Homework
Linear Regression and Correlation: Homework is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Exercise:

Problem: For each situation below, state the independent variable and the dependent variable.

- **a**A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than all other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
- **b**A study is done to determine if the weekly grocery bill changes based on the number of family members.
- **c**Insurance companies base life insurance premiums partially on the age of the applicant.
- **d**Utility bills vary according to power consumption.
- **e**A study is done to determine if a higher education reduces the crime rate in a population.

Solution:

- **a**Independent: Age; Dependent: Fatalities
- **d**Independent: Power Consumption; Dependent: Utility

Note: For any prediction questions, the answers are calculated using the least squares (best fit) line equation cited in the solution.

Exercise:

Problem:

Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows (Source: *http://* *http://www.census.gov/compendia/statab/cats/transportation/motor_vehicle_accidents_and_fatalities.html*):

Age	Number of Driver Deaths per 100,000
16-19	38
20-24	36
25-34	24
35-54	20
55-74	18

Age	Number of Driver Deaths per 100,000
75+	28

- **a**For each age group, pick the midpoint of the interval for the x value. (For the 75+ group, use 80.)
- **b**Using “ages” as the independent variable and “Number of driver deaths per 100,000” as the dependent variable, make a scatter plot of the data.
- **c**Calculate the least squares (best–fit) line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **d**Find the correlation coefficient. Is it significant?
- **e**Pick two ages and find the estimated fatality rates.
- **f**Use the two points in (e) to plot the least squares line on your graph from (b).
- **g**Based on the above data, is there a linear relationship between age of a driver and driver fatality rate?
- **h**What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise:

Problem:

The average number of people in a family that received welfare for various years is given below. (Source: *House Ways and Means Committee, Health and Human Services Department*)

Year	Welfare family size
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

- **a**Using “year” as the independent variable and “welfare family size” as the dependent variable, make a scatter plot of the data.
- **b**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **c**Find the correlation coefficient. Is it significant?
- **d**Pick two years between 1969 and 1991 and find the estimated welfare family sizes.
- **e**Use the two points in (d) to plot the least squares line on your graph from (b).
- **f**Based on the above data, is there a linear relationship between the year and the average number of people in a welfare family?

- **g**Using the least squares line, estimate the welfare family sizes for 1960 and 1995. Does the least squares line give an accurate estimate for those years? Explain why or why not.
- **h**Are there any outliers in the above data?
- **i**What is the estimated average welfare family size for 1986? Does the least squares line give an accurate estimate for that year? Explain why or why not.
- **j**What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

- **b** $\hat{y} = 88.7206 - 0.0432x$
- **c**-0.8533, Yes
- **g**No
- **h**No.
- **i**2.93, Yes
- **j**slope = -0.0432. As the year increases by one, the welfare family size tends to decrease by 0.0432 people.

Exercise:

Problem:

Use the AIDS data from the practice for this section, but this time use the columns “year #” and “# new AIDS deaths in U.S.” Answer all of the questions from the practice again, using the new columns.

Exercise:

Problem:

The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level). (Source: *Microsoft Bookshelf*)

Height (in feet)	Stories
1050	57
428	28
362	26
529	40
790	60
401	22
380	38
1454	110

Height (in feet)	Stories
1127	100
700	46

- **a**Using “stories” as the independent variable and “height” as the dependent variable, make a scatter plot of the data.
- **b**Does it appear from inspection that there is a relationship between the variables?
- **c**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **d**Find the correlation coefficient. Is it significant?
- **e**Find the estimated heights for 32 stories and for 94 stories.
- **f**Use the two points in (e) to plot the least squares line on your graph from (b).
- **g**Based on the above data, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
- **h**Are there any outliers in the above data? If so, which point(s)?
- **i**What is the estimated height of a building with 6 stories? Does the least squares line give an accurate estimate of height? Explain why or why not.
- **j**Based on the least squares line, adding an extra story is predicted to add about how many feet to a building?
- **k**What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

- **b**Yes
- **c** $\hat{y} = 102.4287 + 11.7585x$
- **d**0.9436; yes
- **e**478.70 feet; 1207.73 feet
- **g**Yes
- **h**Yes; $(57,1050)$
- **i**172.98; No
- **j**11.7585 feet
- **k**slope = 11.7585. As the number of stories increases by one, the height of the building tends to increase by 11.7585 feet.

Exercise:

Problem:

Below is the life expectancy for an individual born in the United States in certain years. (Source: *National Center for Health Statistics*)

Year of Birth	Life Expectancy
1930	59.7

Year of Birth	Life Expectancy
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Draw a scatter plot of the ordered pairs.
- **c**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \mathrm{bx}$
- **d**Find the correlation coefficient. Is it significant?
- **e**Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.
- **f**Why aren't the answers to part (e) the values on the above chart that correspond to those years?
- **g**Use the two points in (e) to plot the least squares line on your graph from (b).
- **h**Based on the above data, is there a linear relationship between the year of birth and life expectancy?
- **i**Are there any outliers in the above data?
- **j**Using the least squares line, find the estimated life expectancy for an individual born in 1850. Does the least squares line give an accurate estimate for that year? Explain why or why not.
- **k**What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise:

Problem:

The percent of female wage and salary workers who are paid hourly rates is given below for the years 1979 - 1992. (Source: *Bureau of Labor Statistics, U.S. Dept. of Labor*)

Year	Percent of workers paid hourly rates
1979	61.2
1980	60.7
1981	61.3

Year	Percent of workers paid hourly rates
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

- **a**Using “year” as the independent variable and “percent” as the dependent variable, make a scatter plot of the data.
- **b**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **c**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **d**Find the correlation coefficient. Is it significant?
- **e**Find the estimated percents for 1991 and 1988.
- **f**Use the two points in (e) to plot the least squares line on your graph from (b).
- **g**Based on the above data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
- **h**Are there any outliers in the above data?
- **i**What is the estimated percent for the year 2050? Does the least squares line give an accurate estimate for that year? Explain why or why not?
- **j**What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

- **b**Yes
- **c** $\hat{y} = -266.8863 + 0.1656x$
- **d**0.9448; Yes
- **e**62.8233; 62.3265
- **h** yes; (1987, 62.7)
- **i**72.5937; No
- **j**slope = 0.1656. As the year increases by one, the percent of workers paid hourly rates tends to increase by 0.1656.

Exercise:

Problem:

The maximum discount value of the Entertainment® card for the “Fine Dining” section, Edition 10, for various pages is given below.

Page number	Maximum value ($)
4	16
14	19
25	15
32	17
43	19
57	15
72	16
85	15
90	17

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Draw a scatter plot of the ordered pairs.
- **c**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **d**Find the correlation coefficient. Is it significant?
- **e**Find the estimated maximum values for the restaurants on page 10 and on page 70.
- **f**Use the two points in (e) to plot the least squares line on your graph from (b).
- **g**Does it appear that the restaurants giving the maximum value are placed in the beginning of the "Fine Dining" section? How did you arrive at your answer?
- **h**Suppose that there were 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
- **i**Is the least squares line valid for page 200? Why or why not?
- **j**What is the slope of the least squares (best-fit) line? Interpret the slope.

The next two questions refer to the following data: The cost of a leading liquid laundry detergent in different sizes is given below.

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

Exercise:

Problem:

- **a**Using “size” as the independent variable and “cost” as the dependent variable, make a scatter plot.
- **b**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **c**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **d**Find the correlation coefficient. Is it significant?
- **e**If the laundry detergent were sold in a 40 ounce size, find the estimated cost.
- **f**If the laundry detergent were sold in a 90 ounce size, find the estimated cost.
- **g**Use the two points in (e) and (f) to plot the least squares line on your graph from (a).
- **h**Does it appear that a line is the best way to fit the data? Why or why not?
- **i**Are there any outliers in the above data?
- **j**Is the least squares line valid for predicting what a 300 ounce size of the laundry detergent would cost? Why or why not?
- **k**What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

- **b**Yes
- **c** $\hat{y} = 3.5984 + 0.0371x$
- **d**0.9986; Yes
- **e**$5.08
- **f**$6.93
- **i**No
- **j**Not valid
- **k**slope = 0.0371. As the number of ounces increases by one, the cost of liquid detergent tends to increase by $0.0371 or is predicted to increase by $0.0371 (about 4 cents).

Exercise:

Problem:

- **a**Complete the above table for the cost per ounce of the different sizes.
- **b**Using “Size” as the independent variable and “Cost per ounce” as the dependent variable, make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. Is it significant?
- **f**If the laundry detergent were sold in a 40 ounce size, find the estimated cost per ounce.
- **g**If the laundry detergent were sold in a 90 ounce size, find the estimated cost per ounce.
- **h**Use the two points in (f) and (g) to plot the least squares line on your graph from (b).
- **i**Does it appear that a line is the best way to fit the data? Why or why not?
- **j**Are there any outliers in the above data?
- **k**Is the least squares line valid for predicting what a 300 ounce size of the laundry detergent would cost per ounce? Why or why not?
- **l**What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise:

Problem:

According to flyer by a Prudential Insurance Company representative, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net Taxable Estate ($)	Approximate Probate Fees and Taxes ($)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. Is it significant?
- **f**Find the estimated total cost for a net taxable estate of $1,000,000. Find the cost for $2,500,000.
- **g**Use the two points in (f) to plot the least squares line on your graph from (b).
- **h**Does it appear that a line is the best way to fit the data? Why or why not?
- **i**Are there any outliers in the above data?
- **j**Based on the above, what would be the probate fees and taxes for an estate that does not have any assets?
- **k**What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

- **c**Yes
- **d** $\hat{y} = -337{,}424.6478 + 0.5463x$
- **e**0.9964; Yes
- **f**$208,875.35; $1,028,325.35
- **h**Yes
- **i**No
- **k**slope = 0.5463. As the net taxable estate increases by one dollar, the approximate probate fees and taxes tend to increase by 0.5463 dollars (about 55 cents).

Exercise:

Problem: The following are advertised sale prices of color televisions at Anderson's.

Size (inches)	Sale Price ($)
9	147
20	197
27	297
31	447
35	1177
40	2177
60	2497

- **a** Decide which variable should be the independent variable and which should be the dependent variable.
- **b** Make a scatter plot of the data.
- **c** Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d** Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e** Find the correlation coefficient. Is it significant?
- **f** Find the estimated sale price for a 32 inch television. Find the cost for a 50 inch television.
- **g** Use the two points in (f) to plot the least squares line on your graph from (b).
- **h** Does it appear that a line is the best way to fit the data? Why or why not?
- **i** Are there any outliers in the above data?
- **j** What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise:

Problem: Below are the average heights for American boys. (Source: *Physician's Handbook, 1990*)

Age (years)	Height (cm)
birth	50.8
2	83.8

Age (years)	Height (cm)
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. Is it significant?
- **f**Find the estimated average height for a one year–old. Find the estimated average height for an eleven year–old.
- **g**Use the two points in (f) to plot the least squares line on your graph from (b).
- **h**Does it appear that a line is the best way to fit the data? Why or why not?
- **i**Are there any outliers in the above data?
- **j**Use the least squares line to estimate the average height for a sixty–two year–old man. Do you think that your answer is reasonable? Why or why not?
- **k**What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

- **c**Yes
- **d** $\hat{y} = 65.0876 + 7.0948x$
- **e**0.9761; yes
- **f**72.2 cm; 143.13 cm
- **h**Yes
- **i**No
- **j**505.0 cm; No
- **k**slope = 7.0948. As the age of an American boy increases by one year, the average height tends to increase by 7.0948 cm.

Exercise:

Problem:

The following chart gives the gold medal times for every other Summer Olympics for the women’s 100 meter freestyle (swimming).

Year	Time (seconds)
1912	82.2
1924	72.4
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64
2000	53.8
2008	53.1

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. Is the decrease in times significant?
- **f**Find the estimated gold medal time for 1932. Find the estimated time for 1984.
- **g**Why are the answers from (f) different from the chart values?
- **h**Use the two points in (f) to plot the least squares line on your graph from (b).
- **i**Does it appear that a line is the best way to fit the data? Why or why not?
- **j**Use the least squares line to estimate the gold medal time for the next Summer Olympics. Do you think that your answer is reasonable? Why or why not?

The next three questions use the following state information.

State	# letters in name	Year entered the Union	Rank for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado		1876	38	104,100

State	# letters in name	Year entered the Union	Rank for entering the Union	Area (square miles)
Hawaii		1959	50	10,932
Iowa		1846	29	56,276
Maryland		1788	7	12,407
Missouri		1821	24	69,709
New Jersey		1787	3	8,722
Ohio		1803	17	44,828
South Carolina	13	1788	8	32,008
Utah		1896	45	84,904
Wisconsin		1848	30	65,499

Exercise:

Problem:

We are interested in whether or not the number of letters in a state name depends upon the year the state entered the Union.

- **a**Decide which variable should be the independent variable and which should be the dependent variable.
- **b**Make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **e**Find the correlation coefficient. What does it imply about the significance of the relationship?
- **f**Find the estimated number of letters (to the nearest integer) a state would have if it entered the Union in 1900. Find the estimated number of letters a state would have if it entered the Union in 1940.
- **g**Use the two points in (f) to plot the least squares line on your graph from (b).
- **h**Does it appear that a line is the best way to fit the data? Why or why not?
- **i**Use the least squares line to estimate the number of letters a new state that enters the Union this year would have. Can the least squares line be used to predict it? Why or why not?

Solution:

- **c**No
- **d** $\hat{y} = 47.03 - 0.0216x$
- **e**-0.4280
- **f**6; 5

Exercise:

Problem:

We are interested in whether there is a relationship between the ranking of a state and the area of the state.

- **a**Let rank be the independent variable and area be the dependent variable.
- **b**What do you think the scatter plot will look like? Make a scatter plot of the data.
- **c**Does it appear from inspection that there is a relationship between the variables? Why or why not?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \mathrm{bx}$
- **e**Find the correlation coefficient. What does it imply about the significance of the relationship?
- **f**Find the estimated areas for Alabama and for Colorado. Are they close to the actual areas?
- **g**Use the two points in (f) to plot the least squares line on your graph from (b).
- **h**Does it appear that a line is the best way to fit the data? Why or why not?
- **i**Are there any outliers?
- **j**Use the least squares line to estimate the area of a new state that enters the Union. Can the least squares line be used to predict it? Why or why not?
- **k**Delete "Hawaii" and substitute "Alaska" for it. Alaska is the fortieth state with an area of 656,424 square miles.
- **l**Calculate the new least squares line.
- **m**Find the estimated area for Alabama. Is it closer to the actual area with this new least squares line or with the previous one that included Hawaii? Why do you think that's the case?
- **n**Do you think that, in general, newer states are larger than the original states?

Exercise:

Problem:

We are interested in whether there is a relationship between the rank of a state and the year it entered the Union.

- **a**Let year be the independent variable and rank be the dependent variable.
- **b**What do you think the scatter plot will look like? Make a scatter plot of the data.
- **c**Why must the relationship be positive between the variables?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \mathrm{bx}$
- **e**Find the correlation coefficient. What does it imply about the significance of the relationship?
- **f**Let's say a fifty-first state entered the union. Based upon the least squares line, when should that have occurred?
- **g**Using the least squares line, how many states do we currently have?
- **h**Why isn't the least squares line a good estimator for this year?

Solution:

- **d** $\hat{y} = -480.5845 + 0.2748x$
- **e**0.9553
- **f**1934

Exercise:

Problem:

Below are the percents of the U.S. labor force (excluding self-employed and unemployed) that are members of a union. We are interested in whether the decrease is significant. (Source: *Bureau of Labor Statistics, U.S. Dept. of Labor*)

Year	Percent
1945	35.5
1950	31.5
1960	31.4
1970	27.3
1980	21.9
1993	15.8
2011	11.8

- **a**Let year be the independent variable and percent be the dependent variable.
- **b**What do you think the scatter plot will look like? Make a scatter plot of the data.
- **c**Why will the relationship between the variables be negative?
- **d**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \mathrm{bx}$
- **e**Find the correlation coefficient. What does it imply about the significance of the relationship?
- **f**Based on your answer to (e), do you think that the relationship can be said to be decreasing?
- **g**If the trend continues, when will there no longer be any union members? Do you think that will happen?

The next two questions refer to the following information: The data below reflects the 1991-92 Reunion Class Giving. (Source: *SUNY Albany alumni magazine*)

Class Year	Average Gift	Total Giving
1922	41.67	125
1927	60.75	1,215
1932	83.82	3,772

Class Year	Average Gift	Total Giving
1937	87.84	5,710
1947	88.27	6,003
1952	76.14	5,254
1957	52.29	4,393
1962	57.80	4,451
1972	42.68	18,093
1976	49.39	22,473
1981	46.87	20,997
1986	37.03	12,590

Exercise:

Problem:

We will use the columns “class year” and “total giving” for all questions, unless otherwise stated.

- **a**What do you think the scatter plot will look like? Make a scatter plot of the data.
- **b**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **c**Find the correlation coefficient. What does it imply about the significance of the relationship?
- **d**For the class of 1930, predict the total class gift.
- **e**For the class of 1964, predict the total class gift.
- **f**For the class of 1850, predict the total class gift. Why doesn’t this value make any sense?

Solution:

- **b** $\hat{y} = -569{,}770.2796 + 296.0351x$
- **c**0.8302
- **d**$1577.46
- **e**$11,642.66
- **f**-$22,105.34

Exercise:

Problem:

We will use the columns “class year” and “average gift” for all questions, unless otherwise stated.

- **a**What do you think the scatter plot will look like? Make a scatter plot of the data.
- **b**Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + \text{bx}$
- **c**Find the correlation coefficient. What does it imply about the significance of the relationship?
- **d**For the class of 1930, predict the average class gift.
- **e**For the class of 1964, predict the average class gift.
- **f**For the class of 2010, predict the average class gift. Why doesn’t this value make any sense?

Exercise:

Problem:

We are interested in exploring the relationship between the weight of a vehicle and its fuel efficiency (gasoline mileage). The data in the table show the weights, in pounds, and fuel efficiency, measured in miles per gallon, for a sample of 12 vehicles.

Weight	Fuel Efficiency
2715	24
2570	28
2610	29
2750	38
3000	25
3410	22
3640	20
3700	26
3880	21
3900	18
4060	18
4710	15

- **a**Graph a scatterplot of the data.
- **b**Find the correlation coefficient and determine if it is significant.
- **c**Find the equation of the best fit line.
- **d**Write the sentence that interprets the meaning of the slope of the line in the context of the data.
- **e**What percent of the variation in fuel efficiency is explained by the variation in the weight of the vehicles, using the regression line? (State your answer in a complete sentence in the context of the data.)
- **f**Accurately graph the best fit line on your scatterplot.
- **g**For the vehicle that weights 3000 pounds, find the residual (y-yhat). Does the value predicted by the line underestimate or overestimate the observed data value?
- **h**Identify any outliers, using either the graphical or numerical procedure demonstrated in the textbook.
- **i**The outlier is a hybrid car that runs on gasoline and electric technology, but all other vehicles in the sample have engines that use gasoline only. Explain why it would be appropriate to remove the

outlier from the data in this situation. Remove the outlier from the sample data. Find the new correlation coefficient, coefficient of determination, and best fit line.

- **j**Compare the correlation coefficients and coefficients of determination before and after removing the outlier, and explain in complete sentences what these numbers indicate about how the model has changed.

Solution:

- **b**r = -0.8, significant
- **c**yhat = 48.4-0.00725x
- **d**For every one pound increase in weight, the fuel efficiency tends to decrease (or is predicted to decrease) by 0.00725 miles per gallon. (For every one thousand pounds increase in weight, the fuel efficiency tends to decrease by 7.25 miles per gallon.)
- **e**64% of the variation in fuel efficiency is explained by the variation in weight using the regression line.
- **g**yhat=48.4-0.00725(3000)=26.65 mpg. y-yhat=25-26.65=-1.65. Because yhat=26.5 is greater than y=25, the line overestimates the observed fuel efficiency.
- **h**(2750,38) is the outlier. Be sure you know how to justify it using the requested graphical or numerical methods, not just by guessing.
- **i**yhat = 42.4-0.00578x
- **j**Without outlier, r=-0.885, rsquare=0.76; with outlier, r=-0.8, rsquare=0.64. The new linear model is a better fit, after the outlier is removed from the data, because the new correlation coefficient is farther from 0 and the new coefficient of determination is larger.

Exercise:

Problem:

The four data sets below were created by statistician Francis Anscomb. They show why it is important to examine the scatterplots for your data, in addition to finding the correlation coefficient, in order to evaluate the appropriateness of fitting a linear model.

Set 1			Set 2			Set 3			Set 4	
x	y		x	y		x	y		x	y
10	8.04		10	9.14		10	7.46		8	6.58
8	6.95		8	8.14		8	6.77		8	5.76
13	7.58		13	8.74		13	12.74		8	7.71
9	8.81		9	8.77		9	7.11		8	8.84
11	8.33		11	9.26		11	7.81		8	8.47
14	9.96		14	8.10		14	8.84		8	7.04

6	7.24		6	6.13		6	6.08		8	5.25
4	4.26		4	3.10		4	5.39		19	12.50
12	10.84		12	9.13		12	8.15		8	5.56
7	4.82		7	7.26		7	6.42		8	7.91
5	5.68		5	4.74		5	5.73		8	6.89

a. For each data set, find the least squares regression line and the correlation coefficient. What did you discover about the lines and values of r?

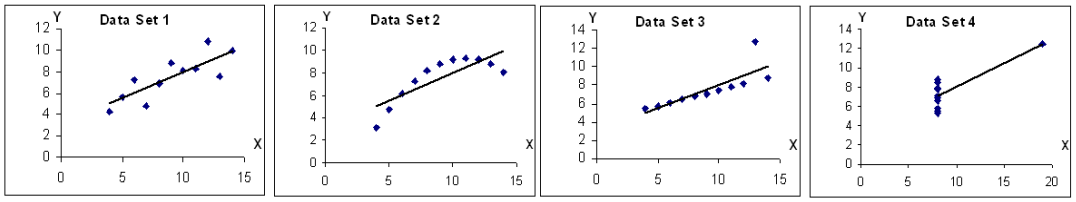
For each data set, create a scatter plot and graph the least squares regression line. Use the graphs to answer the following questions:

- **b**For which data set does it appear that a curve would be a more appropriate model than a line?
- **c**Which data set has an **influential point** (point close to or on the line that greatly influences the best fit line)?
- **d**Which data set has an **outlier** (obviously visible on the scatter plot with best fit line graphed)?
- **e**Which data set appears to be the most appropriate to model using the least squares regression line?

Solution:

a. All four data sets have the same correlation coefficient r=0.816 and the same least squares regression line yhat=3+0.5x

b. Set 2 ; c. Set 4 ; d. Set 3 ; e. Set 1



Try these multiple choice questions

Exercise:

Problem: A correlation coefficient of -0.95 means there is a _____ between the two variables.

- **A**Strong positive correlation
- **B**Weak negative correlation

- **C**Strong negative correlation
- **D**No Correlation

Solution:

C

Exercise:

Problem:

According to the data reported by the New York State Department of Health regarding West Nile Virus **(http://www.health.state.ny.us/nysdoh/westnile/update/update.htm)** for the years 2000-2008, the least squares line equation for the number of reported dead birds (x) versus the number of human West Nile virus cases (y) is $\hat{y} = -10.2638 + 0.0491x$. If the number of dead birds reported in a year is 732, how many human cases of West Nile virus can be expected? $r = 0.5490$

- **A**No prediction can be made.
- **B**19.6
- **C**15
- **D**38.1

Solution:

A

The next three questions refer to the following data: (showing the number of hurricanes by category to directly strike the mainland U.S. each decade) obtained from *www.nhc.noaa.gov/gifs/table6.gif* A major hurricane is one with a strength rating of 3, 4 or 5.

Decade	Total Number of Hurricanes	Number of Major Hurricanes
1941-1950	24	10
1951-1960	17	8
1961-1970	14	6
1971-1980	12	4
1981-1990	15	5
1991-2000	14	5
2001 – 2004	9	3

Exercise:

Problem:

Using only completed decades (1941 – 2000), calculate the least squares line for the number of major hurricanes expected based upon the total number of hurricanes.

- **A** $\hat{y} = -1.67x + 0.5$
- **B** $\hat{y} = 0.5\text{x} - 1.67$
- **C** $\hat{y} = 0.94x - 1.67$
- **D** $\hat{y} = -2\text{x} + 1$

Solution:

B

Exercise:

Problem: The correlation coefficient is 0.942. Is this considered significant? Why or why not?

- **A**No, because 0.942 is greater than the critical value of 0.707
- **B**Yes, because 0.942 is greater than the critical value of 0.707
- **C**No, because 0942 is greater than the critical value of 0.811
- **D**Yes, because 0.942 is greater than the critical value of 0.811

Solution:

D

Exercise:

Problem:

The data for 2001-2004 show 9 hurricanes have hit the mainland United States. The line of best fit predicts 2.83 major hurricanes to hit mainland U.S. Can the least squares line be used to make this prediction?

- **A**No, because 9 lies outside the independent variable values
- **B**Yes, because, in fact, there have been 3 major hurricanes this decade
- **C**No, because 2.83 lies outside the dependent variable values
- **D**Yes, because how else could we predict what is going to happen this decade.

Solution:

A

**Exercises 21 and 22 contributed by Roberta Bloom

Lab 1: Regression (Distance from School)
This module provides a lab of Linear Regression and Correlation as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Class Time:

Names:

Student Learning Outcomes:

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

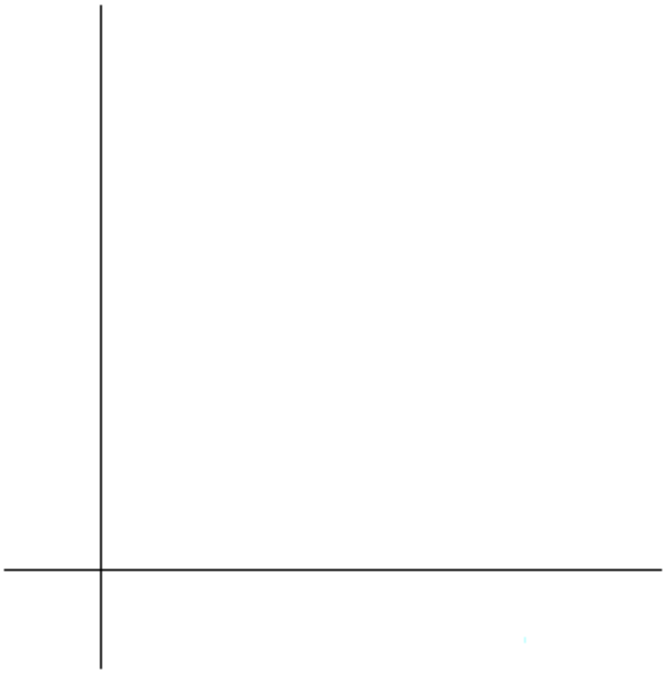
Collect the Data

Use 8 members of your class for the sample. Collect bivariate data (distance an individual lives from school, the cost of supplies for the current term).

1. Complete the table.

Distance from school	Cost of supplies this term

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph “distance” vs. “cost.” Plot the points on the graph. Label both axes with words. Scale both axes.



Analyze the Data

Enter your data into your calculator or computer. Write the linear equation below, rounding to 4 decimal places.

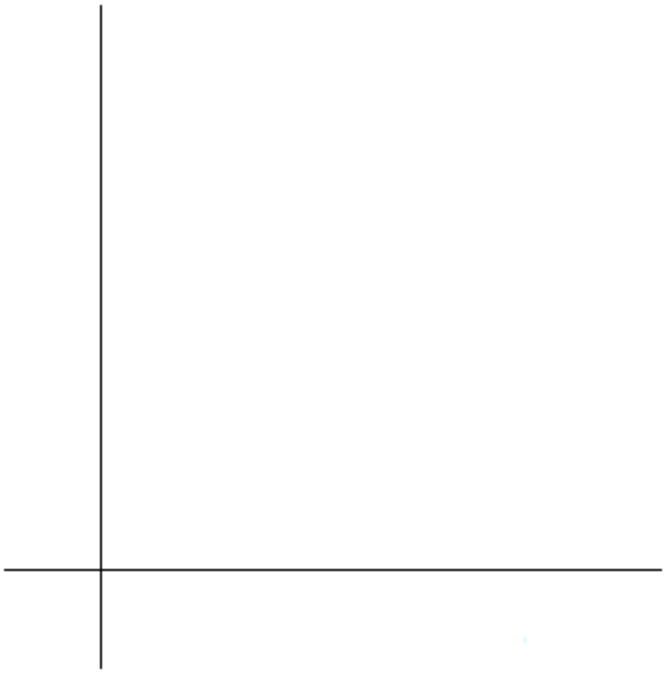
1. Calculate the following:
    - **a** a =
    - **b** b=
    - **c** correlation =
    - **d** n =
    - **e** equation: $\hat{y}$ =

- **f** Is the correlation significant? Why or why not? (Answer in 1-3 complete sentences.)

2. Supply an answer for the following senarios:

    - **a**For a person who lives 8 miles from campus, predict the total cost of supplies this term:
    - **b**For a person who lives 80 miles from campus, predict the total cost of supplies this term:

3. Obtain the graph on your calculator or computer. Sketch the regression line below.



Discussion Questions

1. Answer each with 1-3 complete sentences.

    - **a**Does the line seem to fit the data? Why?

- **b**What does the correlation imply about the relationship between the distance and the cost?

2. Are there any outliers? If so, which point is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

Lab 2: Regression (Textbook Cost)
This module provides a lab of Linear Regression and Correlation as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Class Time:

Names:

Student Learning Outcomes:

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

Collect the Data

Survey 10 textbooks. Collect bivariate data (number of pages in a textbook, the cost of the textbook).

1. Complete the table.

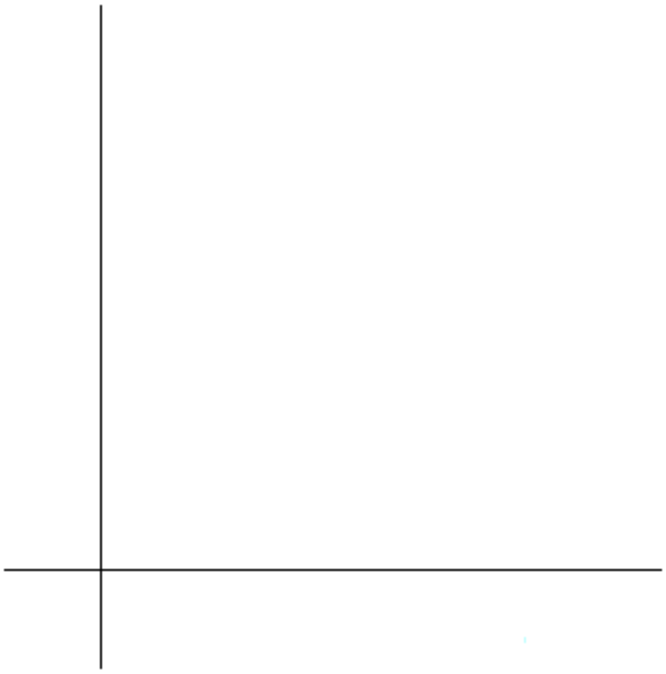
Number of pages	Cost of textbook

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph “distance” vs. “cost.” Plot the points on the graph in "Analyze the Data". Label both axes with words. Scale both axes.

Analyze the Data

Enter your data into your calculator or computer. Write the linear equation below, rounding to 4 decimal places.

1. Calculate the following:
    - **a** a =
    - **b** b =
    - **c** correlation =
    - **d** n =
    - **e** equation: y =
    - **f** Is the correlation significant? Why or why not? (Answer in 1-3 complete sentences.)
2. Supply an answer for the following senarios:
    - **a**For a textbook with 400 pages, predict the cost:
    - **b** For a textbook with 600 pages, predict the cost:
3. Obtain the graph on your calculator or computer. Sketch the regression line below.



Discussion Questions

1. Answer each with 1-3 complete sentences.
    - **a**Does the line seem to fit the data? Why?
    - **b**What does the correlation imply about the relationship between the number of pages and the cost?

2. Are there any outliers? If so, which point(s) is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

Lab 3: Regression (Fuel Efficiency)
This module provides a lab of Linear Regression and Correlation as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Class Time:

Names:

Student Learning Outcomes:

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

Collect the Data

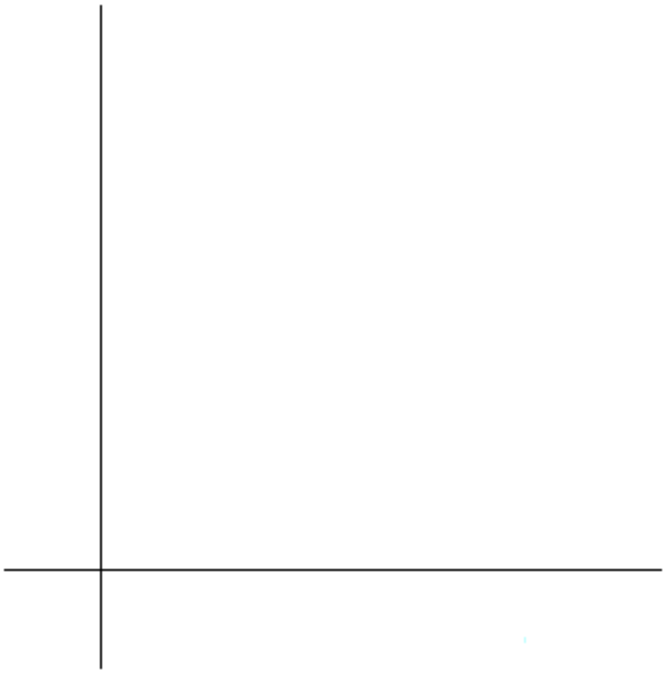
Use the most recent April issue of Consumer Reports. It will give the total fuel efficiency (in miles per gallon) and weight (in pounds) of new model cars with automatic transmissions. We will use this data to determine the relationship, if any, between the fuel efficiency of a car and its weight.

1. Which variable should be the independent variable and which should be the dependent variable? Explain your answer in one or two complete sentences.
2. Using your random number generator, randomly select 20 cars from the list and record their weights and fuel efficiency into the table below.

Weight	**Fuel Efficiency**

Weight	Fuel Efficiency

3. Which variable should be the dependent variable and which should be the independent variable? Why?
4. By hand, do a scatterplot of “weight” vs. “fuel efficiency”. Plot the points on graph paper. Label both axes with words. Scale both axes accurately.



Analyze the Data

Enter your data into your calculator or computer. Write the linear equation below, rounding to 4 decimal places.

1. Calculate the following:
    - **a** a =
    - **b** b =
    - **c** correlation =
    - **d** n =
    - **e** equation: $\hat{y}$ =
2. Obtain the graph of the regression line on your calculator. Sketch the regression line on the same axes as your scatterplot.

Discussion Questions

1. Is the correlation significant? Explain how you determined this in complete sentences.
2. Is the relationship a positive one or a negative one? Explain how you can tell and what this means in terms of weight and fuel efficiency.
3. In one or two complete sentences, what is the practical interpretation of the slope of the least squares line in terms of fuel efficiency and weight?
4. For a car that weighs 4000 pounds, predict its fuel efficiency. Include units.
5. Can we predict the fuel efficiency of a car that weighs 10000 pounds using the least squares line? Explain why or why not.
6. Questions. Answer each in 1 to 3 complete sentences.

    - **a** Does the line seem to fit the data? Why or why not?
    - **b** What does the correlation imply about the relationship between fuel efficiency and weight of a car? Is this what you expected?

7. Are there any outliers? If so, which point is an outlier?

** This lab was designed and contributed by Diane Mathios.

Group Project: Bivariate Data, Linear Regression, and Univariate Data
In this project, students will collect a sample of bivariate data and analyze the infomration. Students will be asked to describe the center and spread of the data, determine the goodness of fit of a linear regression model, and analyze the relationship between the variables.

Student Learning Objectives

- The students will collect a bivariate data sample through the use of appropriate sampling techniques.
- The student will attempt to fit the data to a linear model.
- The student will determine the appropriateness of linear fit of the model.
- The student will analyze and graph univariate data.

Instructions

1. As you complete each task below, check it off. Answer all questions in your introduction or summary.
2. Check your course calendar for intermediate and final due dates.
3. Graphs may be constructed by hand or by computer, unless your instructor informs you otherwise. All graphs must be neat and accurate.
4. All other responses must be done on the computer.
5. Neatness and quality of explanations are used to determine your final grade.

Part I: Bivariate Data

Introduction

- ____State the bivariate data your group is going to study.

> **Note:**Here are two examples, but you may **NOT** use them: height vs. weight and age vs. running distance.

- ____Describe how your group is going to collect the data (for instance, collect data from the web, survey students on campus).
- ____Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random sampling (using a random number generator) sampling. Convenience sampling is **NOT** acceptable.
- ____Conduct your survey. Your number of pairs must be at least 30.
- ____Print out a copy of your data.

Analysis

- ____On a separate sheet of paper construct a scatter plot of the data. Label and scale both axes.
- ____State the least squares line and the correlation coefficient.
- ____On your scatter plot, in a different color, construct the least squares line.
- ____Is the correlation coefficient significant? Explain and show how you determined this.
- ____Interpret the slope of the linear regression line in the context of the data in your project. Relate the explanation to your data, and quantify what the slope tells you.
- ____Does the regression line seem to fit the data? Why or why not? If the data does not seem to be linear, explain if any other model seems to fit the data better.
- ____Are there any outliers? If so, what are they? Show your work in how you used the potential outlier formula in the Linear Regression and Correlation chapter (since you have bivariate data) to determine whether or not any pairs might be outliers.

Part II: Univariate Data

In this section, you will use the data for **ONE** variable only. Pick the variable that is more interesting to analyze. For example: if your

independent variable is sequential data such as year with 30 years and one piece of data per year, your x-values might be 1971, 1972, 1973, 1974, …, 2000. This would not be interesting to analyze. In that case, choose to use the dependent variable to analyze for this part of the project.

- _____Summarize your data in a chart with columns showing data value, frequency, relative frequency, and cumulative relative frequency.
- _____Answer the following, rounded to 2 decimal places:
    - **1**Sample mean =
    - **2**Sample standard deviation =
    - **3**First quartile =
    - **4**Third quartile =
    - **5**Median =
    - **6**70th percentile =
    - **7**Value that is 2 standard deviations above the mean =
    - **8**Value that is 1.5 standard deviations below the mean =
- _____Construct a histogram displaying your data. Group your data into 6 – 10 intervals of equal width. Pick regularly spaced intervals that make sense in relation to your data. For example, do NOT group data by age as 20-26,27-33,34-40,41-47,48-54,55-61 . . . Instead, maybe use age groups 19.5-24.5, 24.5-29.5, . . . or 19.5-29.5, 29.5-39.5, 39.5-49.5, . . .
- _____In complete sentences, describe the shape of your histogram.
- _____Are there any potential outliers? Which values are they? Show your work and calculations as to how you used the potential outlier formula in chapter 2 (since you are now using univariate data) to determine which values might be outliers.
- _____Construct a box plot of your data.
- _____Does the middle 50% of your data appear to be concentrated together or spread out? Explain how you determined this.
- _____Looking at both the histogram AND the box plot, discuss the distribution of your data. For example: how does the spread of the middle 50% of your data compare to the spread of the rest of the data represented in the box plot; how does this correspond to your

description of the shape of the histogram; how does the graphical display show any outliers you may have found; does the histogram show any gaps in the data that are not visible in the box plot; are there any interesting features of your data that you should point out.

Due Dates

- Part I, Intro: _____ (keep a copy for your records)
- Part I, Analysis: _____ (keep a copy for your records)
- Entire Project, typed and stapled: _____

    - ____Cover sheet: names, class time, and name of your study.
    - ____Part I: label the sections “Intro” and “Analysis.”
    - ____Part II:
    - ____Summary page containing several paragraphs written in complete sentences describing the experiment, including what you studied and how you collected your data. The summary page should also include answers to ALL the questions asked above.
    - ____All graphs requested in the project.
    - ____All calculations requested to support questions in data.
    - ____Description: what you learned by doing this project, what challenges you had, how you overcame the challenges.

Note: Include answers to ALL questions asked, even if not explicitly repeated in the items above.